

Roberto Tuberosa · Andreas Graner
Emile Frison *Editors*

Genomics of Plant Genetic Resources

Volume 1. Managing, Sequencing and
Mining Genetic Resources

Genomics of Plant Genetic Resources

Roberto Tuberosa • Andreas Graner • Emile Frison
Editors

Genomics of Plant Genetic Resources

Volume 1. Managing, Sequencing and Mining
Genetic Resources

 Springer

Editors

Roberto Tuberosa
Agricultural Sciences
University of Bologna
Bologna
Italy

Emile Frison
Bioersivity International
Rome
Italy

Andreas Graner
Genebank
IPK
Gatersleben
Sachsen-Anhalt
Germany

ISBN 978-94-007-7571-8 ISBN 978-94-007-7572-5 (eBook)
DOI 10.1007/978-94-007-7572-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013954265

© Springer Science+Business Media Dordrecht 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use. While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

Plant genetic resources constitute the feedstock for the biotechnology and genetic engineering enterprises. Year 2013 marks the 60th anniversary of the discovery of the double helix structure of the DNA molecule. This discovery led to the birth of the new genetics based on genomics. The new genetics is helping to revolutionize plant breeding through both marker-assisted selection and recombinant DNA technology. It is in this context that this informative two-volume book entitled “Genomics of Plant Genetic Resources” edited by Prof. Roberto Tuberosa, Prof. Andrea Graner and Dr. Emile Frison is very timely and welcome.

The book deals with managing plant genetic resources, developing genomics platforms and approaches to investigate plant genetic resources, genome sequencing and crop domestication and mining allelic diversity. The different chapters written by eminent authorities shed much light on problems relating to both theoretical and applied genomics. We owe a deep debt of gratitude to the Editors for this labor of love in the cause of conservation and sustainable use of plant genetic resources. This book shows the pathway for achieving an ever-green revolution in agriculture based on enhancement of productivity in perpetuity without associated ecological harm.

M. S. Swaminathan

Foreword

Who would have believed only two decades ago that plant scientists would have access to nearly the complete genetic code of numerous plant species, including major crop species. The idea of having ready access to whole-genome sequences encompassing 140 million bases seemed like science fiction, let alone having available even larger genomes such as rice at 430 Mb or maize at 2500 Mb. And then proceeding to identify variation at the DNA level well beyond what was anticipated, such as the 2.6 SNPs (Single Nucleotide Polymorphisms) per kb in rice. Also produced at an unprecedented rate were literally hundreds of thousands of insert strains, allowing the association of sequences and traits. Who would have believed only a decade ago that we would be capable of analyzing the expression of genes across the whole genome and matching that profile with traits of interest. And now the area of metabolomics is allowing even more meaningful explanations of the biochemical and genetic pathways underlying important traits.

This book brings all of these advances in genomics to the forefront and prepares the plant scientist for the decade ahead. Important technologies are discussed such as association mapping, simulation modeling, and development of appropriate populations including advanced backcrosses and introgression-lines for incorporating traits into useful genetic materials. Such approaches are facilitating the identification of traits that are not obvious simply from observing the plant phenotype, and they provide ways to extract new and useful traits from wild related species. Comparing the genomic information across broadly-related species has generated important evolutionary information. In addition, the common occurrence of duplicated segments recognized in such studies may lead to information fundamental to plant performance.

Methods for the identification of genes underlying traits are improving every day. The association between allelic variation in a candidate gene and a trait is leading to a much greater understanding of the genetic control of traits. Numerous transcription factors and even non-coding sequences are being implicated as the basis of important genetic variation. Forward and reverse genetics are both found to be very useful in making these gene-trait associations.

The tremendous expansion of genomic analytical approaches along with efforts to reduce the cost, together with appropriate statistical designs and analyses, is making it easier and easier to use the ever-increasing sequence information to identify useful genes and gene families. This body of knowledge in plant genomics and its myriad of applications are nicely reflected in this book.

Ronald L. Phillips

Preface

This two-volume book collects 48 manuscripts that present a timely state-of-the-art view on how genomics of plant genetic resources contributes to improve our capacity to characterize and harness natural and artificially induced variation in order to select better cultivars while providing consumers with high-quality and nutritious food. In the past decade, the appreciation of the value of biodiversity has grown steadily, mainly due to the increased awareness of the pivotal role of plant genetic resources for securing the future supply of plant-derived products in the quantity required to meet the burgeoning needs of mankind. The remarkable progress made possible with the deployment of genomics and sequencing platforms has considerably accelerated the pace of gene discovery, the identification of novel, valuable alleles at target loci and their exploitation in breeding programs via marker-assisted selection or other molecular means. Clearly, a better understanding of the genetic make-up and functional variability underpinning the productivity of crops and their adaptation to abiotic and biotic constraints offers unprecedented opportunities for highly targeted approaches while shedding light on the molecular functions that govern such variability.

Meeting the challenges posed by climate change and the future needs of mankind for plant-derived products will require a quantum leap in productivity of the handful of species that provide the staple for our diet and existence. This quantum leap will only be possible through a more effective integration of genomics research with extant breeding programs. As we anticipate a further reduction in the cost of genotyping/sequencing, the exploitation of still largely untapped samples of wild germplasm stored in gene banks will become instrumental for the success of breeding programs. Importantly, the new selection paradigm ushered in by genomics greatly facilitates mining the genetic richness present in orphan crops and underutilized species, previously less readily accessible via conventional approaches.

The unifying picture that emerges from this book unequivocally shows the pivotal role played by genomics to characterize germplasm collections, mine genebanks, elucidate gene function, identify agronomically superior alleles and, ultimately, release improved cultivars. For each of these objectives, the book presents compelling case studies and examples; additional case studies are provided by the references of each chapter.

We hope that this book will provide a helpful reference to students, young researchers, crop specialists and breeders interested in a more effective characterization and utilization of plant genetic resources. In particular, we hope that reading of this book will encourage students and young scientists to pursue a career focused on the study of plant genetic resources and join forces with those already engaged in this challenging and equally fascinating field of science.

We wish to thank all the authors for their timely contributions that have made this book possible. We also thank all those who have contributed to the editing of this book. Last but not least, we wish to thank the policy makers and funding agencies that provide the funds required to collect, conserve, characterize and harness the allelic richness of plant genetic resources.

Roberto Tuberosa
Andreas Graner
Emile Frison

Contents

Part I Managing Genetic Resources

- 1 Building a Global Plant Genetic Resources System** 3
Emile Frison and Nicole Demers
- 2 Genomic Approaches and Intellectual Property Protection for Variety Release: A Perspective from the Private Sector** 27
J. Stephen C. Smith, Elizabeth S. Jones, Barry K. Nelson,
Debra S. Phillips and Robin A. Wineland
- 3 The Use of Molecular Marker Data to Assist in the Determination of Essentially Derived Varieties** 49
J. Stephen C. Smith, Elizabeth S. Jones and Barry K. Nelson
- 4 Application of Molecular Markers in Spatial Analysis to Optimize *In Situ* Conservation of Plant Genetic Resources** 67
Maarten van Zonneveld, Ian Dawson, Evert Thomas, Xavier Scheldeman,
Jacob van Etten, Judy Loo and José I Hormaza
- 5 Historical and Prospective Applications of ‘Quantitative Genomics’ in Utilising Germplasm Resources** 93
Adrian Hathorn and Scott C. Chapman

Part II Platforms and Approaches to Investigate Plant Genetic Resources

- 6 High-throughput SNP Profiling of Genetic Resources in Crop Plants Using Genotyping Arrays** 113
Martin W. Ganai, Ralf Wieseke, Hartmut Luerksen, Gregor Durstewitz,
Eva-Maria Graner, Joerg Plieske and Andreas Polley
- 7 Paleogenomics as a Guide for Traits Improvement** 131
Jérôme Salse

8 Non-invasive Phenotyping Methodologies Enable the Accurate Characterization of Growth and Performance of Shoots and Roots . . .	173
Marcus Jansen, Francisco Pinto, Kerstin A. Nagel, Dagmar van Dusschoten, Fabio Fiorani, Uwe Rascher, Heike U. Schneider, Achim Walter and Ulrich Schurr	
9 Association Mapping of Genetic Resources: Achievements and Future Perspectives	207
Sivakumar Sukumaran and Jianming Yu	
10 Exploiting Barley Genetic Resources for Genome Wide Association Scans (GWAS)	237
Robbie Waugh, Andrew J. Flavell, Joanne Russell, William (Bill) Thomas, Luke Ramsay and Jordi Comadran	
11 Production and Molecular Cytogenetic Identification of Wheat-Alien Hybrids and Introgression Lines	255
Márta Molnár-Láng, István Molnár, Éva Szakács, Gabriella Linc and Zoltán Bedö	
12 Radiation Hybrids: A valuable Tool for Genetic, Genomic and Functional Analysis of Plant Genomes	285
Ajay Kumar, Filippo M. Bassi, Monika K. Michalak de Jimenez, Farhad Ghavami, Mona Mazaheri, Kristin Simons, Muhammad J. Iqbal, Mohamed Mergoum, Shahryar F. Kianian and Penny M. A. Kianian	
13 FISHIS: A New Way in Chromosome Flow Sorting Makes Complex Genomes More Accessible	319
Sergio Lucretti, Debora Giorgi, Anna Farina and Valentina Grosso	
14 Mining Genetic Resources <i>via</i> Ecotilling	349
Bradley J. Till	
Part III Genome Sequencing and Crop Domestication	
15 Next Generation Sequencing and Germplasm Resources	369
Paul Visendi, Jacqueline Batley and David Edwards	
16 Advances in Sequencing the Barley Genome	391
Nils Stein and Burkhard Steuernagel	
17 The Wheat Black Jack: Advances Towards Sequencing the 21 Chromosomes of Bread Wheat	405
Frédéric Choulet, Mario Caccamo, Jonathan Wright, Michael Alaux, Hana Šimková, Jan Šafář, Philippe Leroy, Jaroslav Doležel, Jane Rogers, Kellye Eversole and Catherine Feuillet	

18 Wheat Domestication: Key to Agricultural Revolutions Past and Future	439
Justin D. Faris	
19 Molecular Evidence for Soybean Domestication	465
Kyujung Van, Moon Young Kim, Jin Hee Shin, Kyung Do Kim, Yeong-Ho Lee and Suk-Ha Lee	
20 Genomics of Origin, Domestication and Evolution of <i>Phaseolus vulgaris</i>	483
Elisa Bellucci, Elena Bitocchi, Domenico Rau, Monica Rodriguez, Eleonora Biagetti, Alessandro Giardini, Giovanna Attene, Laura Nanni, Roberto Papa	
Part IV Mining Allelic Diversity	
21 Advances in Nicotiana Genetic and “Omics” Resources	511
James N. D. Battey, Nicolas Sierro, Nicolas Bakaheer and Nikolai V. Ivanov	
22 Mining SNPs and Linkage Analysis in <i>Cynara cardunculus</i>	533
Sergio Lanteri, Alberto Acquadro, Davide Scaglione and Ezio Portis	
23 Genetic Diversity Assessment in European Cynara Collections	559
Mario Augusto Pagnotta and Arshiya Noorani	
24 Analysis and Exploitation of Cereal Genomes with the Aid of <i>Brachypodium</i>	585
Hikmet Budak, Pilar Hernandez and Alan H. Schulman	
25 Mining Natural Variation for Maize Improvement: Selection on Phenotypes and Genes	615
Shilpa Sood, Sherry Flint-Garcia, Martha C. Willcox and James B. Holland	
26 Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward	651
Dario Grattapaglia	
27 Genetic Diversity in the Grapevine Germplasm	683
Federica Cattonaro, Raffaele Testolin, Simone Scalabrin, Michele Morgante and Gabriele Di Gaspero	
Index	705

Contributors

Alberto Acquadro University of Torino, DISAFA, Grugliasco, Italy

Michael Alaux INRA Centre de Versailles-Grignon, Unité de Recherche en Génomique-Info, Versailles, France

Giovanna Attene Dipartimento di Agraria, Università degli Studi di Sassari, Sassari, Italy

Nicolas Bakaher Philip Morris International R&D, Philip Morris Products SA, Neuchâtel, Switzerland

Filippo M. Bassi Durum Wheat Breeding, ICARDA, Rabat, Morocco

Jacqueline Batley University of Queensland, School of Agriculture and Food Sciences, Australia

James N. D. Battey Philip Morris International R&D, Philip Morris Products SA, Neuchâtel, Switzerland

Zoltán Bedő Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Hungary

Elisa Bellucci Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, Ancona, Italy

Eleonora Biagetti Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, via Brecce Bianche, Italy

Elena Bitocchi Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, Ancona, Italy

Hikmet Budak Biological Sciences and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli, Tuzla-Istanbul, Turkey

Mario Caccamo The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, UK

Federica Cattonaro Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, Udine, Italy

Scott C. Chapman CSIRO Plant Industry, Queensland Bioscience Precinct, QLD, Australia

Frédéric Choulet Genetics, Diversity and Ecophysiology of Cereals, INRA Joint Research Unit 1095 Genetics, Clermont-Ferrand, France

Genetics, Diversity and Ecophysiology of Cereals, University Blaise Pascal Joint Research Unit 1095 Genetics, Clermont-Ferrand, France

Jordi Comadran The James Hutton Institute, Invergowrie, Dundee, Scotland

Ian Dawson The World Agroforestry Centre, Headquarters, Nairobi, Kenya

Nicole Demers Bioversity International, Rome, Italy

Gabriele Di Gaspero Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, Udine, Italy

Dipartimento di Scienze Agrarie e Ambientali, University of Udine, Udine, Italy

Jaroslav Doležel Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic

Gregor Durstewitz TraitGenetics GmbH, Gatersleben, Germany

Dagmar van Dusschoten Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

David Edwards University of Queensland, School of Agriculture and Food Sciences, Brisbane, Australia

Jacob van Etten Bioversity International, Turrialba office, Costa Rica

Kellye Eversole International Wheat Genome Sequencing Consortium, Eversole Associates, Bethesda, USA

Anna Farina ENEA—Italian National Agency for New Technologies, Energy and the Environment, Casaccia Research Center, Rome, Italy

Justin D. Faris USDA-Agricultural Research Service, Cereal Crops Research Unit, Fargo, USA

Catherine Feuillet Genetics, Diversity and Ecophysiology of Cereals, INRA Joint Research Unit 1095 Genetics, Clermont-Ferrand, France

Genetics, Diversity and Ecophysiology of Cereals, University Blaise Pascal Joint Research Unit 1095 Genetics, Clermont-Ferrand, France

Fabio Fiorani Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Andrew J. Flavell Division of Plant Sciences, The University of Dundee at JHI, Dundee, Scotland

Sherry Flint-Garcia USDA-ARS Plant Genetics Research Unit, Columbia, USA
Division of Plant Sciences, University of Missouri, Columbia, USA

Emile Frison Bioversity International, Rome, Italy

Martin W. Ganai TraitGenetics GmbH, Gatersleben, Germany

Farhad Ghavami Department of Plant Pathology, University of Minnesota, St. Paul, USA

Alessandro Giardini Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, Ancona, Italy

Debora Giorgi ENEA—Italian National Agency for New Technologies, Energy and the Environment, Casaccia Research Center, Rome, Italy

Eva-Maria Graner TraitGenetics GmbH, Gatersleben, Germany

Dario Grattapaglia EMBRAPA Genetic Resources and Biotechnology—EPqB, Brasilia, Brazil

Universidade Catolica de Brasília- SGAN, Brasilia, Brazil

Valentina Grosso ENEA—Italian National Agency for New Technologies, Energy and the Environment, Casaccia Research Center, Rome, Italy

Adrian Hathorn CSIRO Plant Industry, Queensland Bioscience Precinct, QLD, Australia

Pilar Hernandez Institute for Sustainable Agriculture (IAS-CSIC), Córdoba, Spain

James B. Holland Department of Crop Science, North Carolina State University, Raleigh, USA

USDA-ARS Plant Science Research Unit, Raleigh, USA

José I Hormaza Instituto de Hortofruticultura Subtropical y Mediterránea, (IHSMUMA-CSIC), Estación Experimental La Mayora, Algarrobo-Costa, Spain

Muhammad J. Iqbal Department of Plant Sciences, North Dakota State University, Fargo, USA

Nikolai V. Ivanov Philip Morris International R&D, Philip Morris Products SA, Neuchâtel, Switzerland

Marcus Jansen Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Elizabeth S. Jones Syngenta Biotechnology, Inc., Raleigh, North Carolina

Penny M. A. Kianian Department of Horticultural Science, University of Minnesota, St. Paul, MN 55108, USA

Shahryar F. Kianian USDA-ARS Cereal Disease Laboratory, University of Minnesota, St. Paul, USA

Kyung Do Kim Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Korea

Moon Young Kim Plant Genomics and Breeding Institute, Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, Seoul, Korea

Ajay Kumar Department of Plant Sciences, North Dakota State University, Fargo, USA

Sergio Lanteri University of Torino, DISAFA, Grugliasco, Italy

Suk-Ha Lee Plant Genomics and Breeding Institute, Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea

Yeong-Ho Lee Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea

Philippe Leroy Genetics, Diversity and Ecophysiology of Cereals, INRA Joint Research Unit 1095 Genetics, Clermont-Ferrand, France

Genetics, Diversity and Ecophysiology of Cereals, University Blaise Pascal Joint Research Unit 1095 Genetics, Clermont-Ferrand, France

Gabriella Linc Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Martonvásár, Hungary

Judy Loo Bioersivity International, Headquarters, Rome, Italy

Sergio Lucretti ENEA—Italian National Agency for New Technologies, Energy and the Environment, Casaccia Research Center, Rome, Italy

Hartmut Luerksen TraitGenetics GmbH, Gatersleben, Germany

Mona Mazaheri Department of Plant Sciences, North Dakota State University, Fargo, USA

Mohamed Mergoum Department of Plant Sciences, North Dakota State University, Fargo, USA

Monika K. Michalak de Jimenez Department of Plant Sciences, North Dakota State University, Fargo, USA

István Molnár Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Martonvásár, Hungary

Márta Molnár-Láng Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Martonvásár, Hungary

Michele Morgante Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, 33100 Udine, Italy

Dipartimento di Scienze Agrarie e Ambientali, University of Udine, via delle scienze 208, 33100 Udine, Italy

Kerstin A. Nagel Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Laura Nanni Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, via Brecce Bianche, 60131 Ancona, Italy

Barry K. Nelson DuPont Pioneer, Johnston, Iowa, USA

Arshiya Noorani Plant Production and Protection Division, FAO, Viale delle Terme di Caracalla, Rome 00153, Italy

Mario Augusto Pagnotta Department of Science and Technologies for Agriculture, Forestry, Nature and Energy (DAFNE), University of Tuscia, Viterbo, Italy

Roberto Papa Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche, via Brecce Bianche, 60131 Ancona, Italy

Consiglio per la Ricerca e Sperimentazione in Agricoltura, Cereal Research Centre (CRA-CER), Foggia, Italy

Debora S. Phillips DuPont Pioneer, Johnston, Iowa, USA

Francisco Pinto Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Joerg Plieske TraitGenetics GmbH, Gatersleben, Germany

Andreas Polley TraitGenetics GmbH, Gatersleben, Germany

Ezio Portis University of Torino, DISAFA, Grugliasco, Italy

Luke Ramsay The James Hutton Institute, Dundee, Scotland

Uwe Rascher Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Domenico Rau Dipartimento di Agraria, Università degli Studi di Sassari, 07100 Sassari, Italy

Monica Rodriguez Dipartimento di Agraria, Università degli Studi di Sassari, 07100 Sassari, Italy

Jane Rogers The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, UK

Joanne Russell The James Hutton Institute, Dundee, Scotland

Jan Šafář Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic

Jérôme Salse INRA, UMR 1095 ‘Génétique, Diversité et Ecophysiologie des Céréales’, Laboratory ‘Plant Paleogenomics for Traits Improvement’, Clermont Ferrand, France

Davide Scaglione University of Torino, DISAFA, Grugliasco, Italy

Simone Scalabrin Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, 33100 Udine, Italy

Xavier Scheldeman Bioersity International, Regional Office for the Americas, Cali, Colombia

Ghent University, Faculty of Bioscience Engineering, Gent, Belgium

Heike U. Schneider Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Alan H. Schulman Department of Biotechnology and Food Research, MTT Agrifood Research, Jokioinen, Finland

Institute of Biotechnology, Viikki Biocenter, University of Helsinki, Helsinki, Finland

Ulrich Schurr Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Jin Hee Shin Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea

Nicolas Sierro Philip Morris International R&D, Philip Morris Products SA, Neuchâtel, Switzerland

Hana Šimková Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Olomouc, Czech Republic

Kristin Simons Department of Plant Sciences, North Dakota State University, Fargo, USA

J. Stephen C. Smith DuPont Pioneer, Johnston, Iowa, USA

Shilpa Sood Department of Crop Science, North Carolina State University, Raleigh, USA

Nils Stein Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Burkhard Steuernagel Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Sivakumar Sukumaran International Maize and Wheat Improvement Center (CIMMYT), Mexico, Mexico

Éva Szakács Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences, Martonvásár, Hungary

Raffaele Testolin Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli, 33100 Udine, Italy

Dipartimento di Scienze Agrarie e Ambientali, University of Udine, via delle scienze 208, 33100 Udine, Italy

Evert Thomas Bioversity International, Regional Office for the Americas, Cali, Colombia

Ghent University, Faculty of Bioscience Engineering, Gent, Belgium

William (Bill) Thomas The James Hutton Institute, Dundee, Scotland

Bradley J. Till Plant Breeding and Genetics Laboratory, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, International Atomic Energy Agency, Vienna International Centre, Vienna, Austria

Kyujung Van Department of Plant Science and Research Institute for Agriculture and Life Sciences, Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea

Paul Visendi University of Queensland, School of Agriculture and Food Sciences, Brisbane, Australia

Achim Walter Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich, Germany

Institute of Agricultural Sciences, ETH Zürich, Zürich, Switzerland

Robbie Waugh The James Hutton Institute, Dundee, Scotland

Ralf Wieseke TraitGenetics GmbH, Gatersleben, Germany

Martha C. Willcox Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT), Texcoco, México

Robin A. Wineland DuPont Pioneer, Johnston, Iowa, USA

Jonathan Wright The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, UK

Jianming Yu Department of Agronomy, Iowa State University, Ames, USA

Maarten van Zonneveld Bioversity International, Turrialba office, Costa Rica

Part I
Managing Genetic Resources

Chapter 1

Building a Global Plant Genetic Resources System

Emile Frison and Nicole Demers

Contents

1.1	Introduction: Global Situation	5
1.1.1	World Population, Hunger and Malnutrition	5
1.1.2	Food Production Situation	6
1.1.3	Climate Change	6
1.2	How can Agriculture Meet Those Challenges?	9
1.2.1	Changes Needed in Agricultural Systems	9
1.2.2	Use of Genetic Diversity and Agrobiodiversity	9
1.2.3	Genetic Resources in Details	9
1.3	A Global System for the Conservation and Sustainable Use of Plant Genetic Resources	11
1.3.1	Today's Situation	11
1.3.2	An Evolving Global System: Some Historical References	12
1.3.3	Elements of a Global System	12
1.4	Conclusion	22
	References	23

Abstract The greatest challenge facing humanity today is to feed tomorrow's population of more than 9 billion people. Production has to increase by about 70 % with the additional uncertainties associated with climate change, against a background of less land and less water being available for agriculture. More than ever before, this will require the wise use of plant genetic resources. Scientific advances such as high-throughput sequencing, marker assisted selection and direct manipulation of the genome have allowed breeders to identify traits and incorporate them into improved varieties more efficiently and more rapidly. The problem is that the genetic resources that are the foundation of these efforts are not being managed effectively. *Ex-situ* collections are currently scattered across roughly 1750 genebanks, many of which are in poor physical condition and which continue to be degraded as a result

E. Frison (✉) · N. Demers
Bioversity International, Via dei Tre Denari 472/a,
Maccarese (Fiumicino), 00057 Rome, Italy
e-mail: e.frison@cgiar.org

N. Demers
e-mail: nicole.demers@cgiar.org

of insufficient and insecure funding. Many of the accessions are duplicates, which is a waste of precious resources. There is little publicly available information about the accessions. Crop wild relatives, which are so important for resistance to biotic and abiotic stresses, are poorly represented in genebanks and in any case need also to be conserved in the wild so that they can continue to evolve in response to those stresses. There is an urgent need to address all these issues by building an effective global system for the conservation and use of plant genetic resources. It will require close collaboration and partnership to ensure efficiency, which in turn will require a commitment to a global system of access and benefit sharing as foreseen by the International Treaty on Plant Genetic Resources for Food and Agriculture. It will require secure and sustainable funding so that we do not have to go through this process again every few decades. And it will require a global information system that guarantees access to much more useful information as well as to the accessions themselves. The challenges are many and complex. As the paper will show, we have the means to meet them, if we engage strongly now, and if we do not, we have little hope of feeding the future population adequately.

Keywords Plant genetic resources · Conservation and sustainable use · Global system · Climate change · Food security · Nutrition

Abbreviations

AEGIS	A European Genebank Integrated System
AVRDC	World Vegetable Centre
BGRI	Borlaug Global Rust Initiative
CAAS	Chinese Academy of Agricultural Sciences
CATIE	Tropical Agricultural Research and Higher Education Center
CBD	Convention on Biological Diversity
CGIAR	Consultative Group on International Agricultural Research
CGRFA	Commission on Genetic Resources for Food and Agriculture
CIMMYT	International Center for Maize and Wheat Improvement
CIP	International Potato Center
CO ₂	Carbon Dioxide
COP-10	Tenth Conference of the Parties of the CBD
CRP	CGIAR Research Programme
CWR	Crop Wild Relatives
EAPGRIN	East African Plant Genetic Resources Network
ECPGR	European Cooperative Programme for Genetic Resources
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EURISCO	European Internet Search Catalogue
FAO	Food and Agriculture Organization of the United Nations
GCDT	Global Crop Diversity Trust
GCP	CGIAR Generation Challenge Programme
GEF	Global Environmental Facility

GHG	Greenhouse Gas
GILB	Global Initiative on Late Blight
GIS	Geographical Information System
GPA	Global Plan of Action on Plant Genetic Resources for Food and Agriculture
GRI	Global Rust Initiative
GRIN	Germplasm Resources Information Network
IBPGR	International Board for Plant Genetic Resources
ICARDA	International Center for Agricultural Research in the Dry Areas
ICWG-GR	CGIAR Inter-Centre Working Group on Genetic Resources
ILRI	International Livestock Research Institute
IPCC	Intergovernmental Panel on Climate Change
ITPGRFA	International Treaty on Plant Genetic Resources for Food and Agriculture
IUCN	International Union for Conservation of Nature
LB	Late Blight
NARS	National Agricultural Research System
NBPGR-India	Indian National Bureau of Plant Genetic Resources
NGRP	National Genetic Resources Program
NordGen	Nordic Genetic Resources Centre
NUS	Neglected and Underutilized Species
PGRFA	Plant Genetic Resources for Food and Agriculture
SANPGR	South Asia Network on Plant Genetic Resources
SGRP	CGIAR System-wide Genetic Resources Programme
SGSV	Svalbard Global Seed Vault
SINGER	CGIAR System-wide Information Network for Genetic Resources
SPGRC	Southern African Development Community Plant Genetic Resources Centre
SRES	Special Report on Emissions Scenarios
SMTA	Standard Material Transfer Agreement
UNEP	United Nations Environmental Programme
USA	United States of America
USDA	United States Department of Agriculture
USDA-ARS	Agricultural Research Service of the USDA

1.1 Introduction: Global Situation

1.1.1 World Population, Hunger and Malnutrition

The greatest challenge facing humanity today is to feed tomorrow's population of more than 9 billion people. In fact, estimates indicate that the world population will reach between 7.5 and 11 billion people by 2050, depending on the expected average number of children per woman (IAASTD 2009). The food price crisis of

2007–2008 caused an increase in the number of hungry people worldwide: of the 923 million in 2007 to 963 million in 2008 (FAO 2008) the world hungry increased at 1,023 million people in 2009 (FAO 2010a). Of this total of 1,023 million, 642 million are in Asia and the Pacific (almost two-thirds), 265 million are in Sub-Saharan Africa (a bit less than one third), 53 million are in Latin America and the Caribbean, 42 million are in the Near East and North Africa and 15 million are in developed countries (FAO 2009). Moreover, malnutrition, the so-called “hidden hunger” caused by micronutrient deficiencies which undermine normal growth and development, health and productivity, is affecting at least 2 billion people worldwide, mostly women and children (Micronutrient Initiative 2009).

1.1.2 Food Production Situation

With the current global economic crisis, the food price crisis of 2007–2008 and climate change, reversing this world hunger trend will be a significant challenge. Since the beginning of the 80’s, official governmental development aid devoted to agriculture has declined and, in 2007, was 37 % lower than in 1988 (FAO 2009). Growth rates of yields for major cereals (wheat and rice) in developing countries have also declined substantially. Investment in agriculture is needed to rejuvenate cereal yield growth rates (FAO 2009).

Global food production will need to increase by 70 % (and nearly to 100 % in developing countries, relative to 2009 level (FAO 2011a)) to cope with a 40 % increase in world population and to raise average food consumption to 3130 kcal per person per day by 2050. This translates into an additional billion tonnes of cereals and 200 million tonnes of meat to be produced annually by 2050 (as compared with production in 2005/07) (Bruinsma 2009). Bruinsma (2009) estimates that ninety percent (and 80 % in developing countries) of the growth in crop production would be a result of higher yields and increased cropping intensity, with the remainder coming from land expansion. The world as a whole produces or could produce enough food for all but land and water, even if globally more than sufficient, are very unevenly distributed and those countries that need to produce more in the future won’t be favoured: the average availability of cultivated land per capita in low-income countries is less than half that of high-income countries, and the suitability of cultivated land for cropping is generally lower. Some countries with rapidly growing demand for food are also those that face high levels of land or water scarcity (Bruinsma 2009, FAO 2011a).

1.1.3 Climate Change

From 1990 to 2005, temperature increases of about 0.2 °C per decade have been observed, strengthening confidence in near-term projections which predict a warming

of about 0.2 °C per decade for the next two decades (from 2007 to 2027) for a range of SRES emissions scenarios¹ (IPCC 2007). Even if the concentrations of all greenhouse gas (GHGs) and aerosols are kept constant at year 2000 levels, a further warming of about 0.1 °C per decade would be expected. Afterwards, temperature projections increasingly depend on specific SRES emissions scenarios, reaching up to 2 °C increase with the worst scenario in 2060 (IPCC 2007). The Intergovernmental Panel on Climate Change (IPCC) and others make the list of some of the consequences that climate change could have and which could affect agricultural production:

- **Species extinction risks:** 30 % of animal and plant species could risk extinction if temperature rise reaches or exceeds 2.5 °C's prediction (IPCC 2007);
- **More extreme weather events:** Altered frequencies and intensities of extreme weather (more heat waves, heavy precipitations, drought, tropical cyclone, etc.) together with sea level rise, are expected to have mostly adverse effects on natural and human systems (IPCC 2007);
- **Changes in ecosystem structure and function:** For increases in global average temperature exceeding 1.5 to 2.5 °C and in concomitant atmospheric carbon dioxide (CO₂) concentrations, there are projected to be major changes in ecosystem structure and function, species' ecological interactions and shifts in species' geographical ranges, with predominantly negative consequences for biodiversity and ecosystem goods and services, e.g. water and food supply (IPCC 2007)
- **Season unpredictability, changes in growing conditions and crop productivity:** Thomas et al. (2007) made specific analysis to distinguish between variability as an expected climate phenomenon and increased variability linked with unpredictability. The main perceptions of change are in the increased variability and uncertainty of specific climate parameters: rains starting later, shorter wet seasons characterised by little but intense rainfall, periodicity of drought and more

¹ "SRES scenarios refer to the scenarios described in the IPCC Special Report on Emissions Scenarios (SRES 2000). The SRES scenarios are grouped into four scenario families (A1, A2, B1 and B2) that explore alternative development pathways, covering a wide range of demographic, economic and technological driving forces and resulting GHG emissions. The SRES scenarios do not include additional climate policies above current ones. The emissions projections are widely used in the assessments of future climate change, and their underlying assumptions with respect to socio-economic, demographic and technological change serve as inputs to many recent climate change vulnerability and impact assessments. The A1 storyline assumes a world of very rapid economic growth, a global population that peaks in mid-century and rapid introduction of new and more efficient technologies. A1 is divided into three groups that describe alternative directions of technological change: fossil intensive (A1FI), non-fossil energy resources (A1T) and a balance across all sources (A1B). B1 describes a convergent world, with the same global population as A1, but with more rapid changes in economic structures toward a service and information economy. B2 describes a world with intermediate population and economic growth, emphasising local solutions to economic, social, and environmental sustainability. A2 describes a very heterogeneous world with high population growth, slow economic development and slow technological change. No likelihood has been attached to any of the SRES scenarios" (IPCC, 2007).

intense heat in the summer were expressed as the key concerns about observed changes in, and unpredictability of, patterns of seasonality. Climate change will cause shifts in areas suitable for cultivation of a wide range of crops. Most detrimentally affected in terms of reduction of suitable areas for a range of crops will be sub-Saharan Africa and the Caribbean. On the contrary, Europe and North America will see an increase in area suitable for cultivation. The former areas are also the ones with the least capacity to cope, as the latter are the regions with the greatest capacity to manage climate change impacts (Lane and Jarvis 2007). As mentioned earlier, marginal desert lands that already pose a challenge to the world's farmers will grow as warmer temperatures fuel desertification. Increasing water scarcity means that we will need to learn to grow more crops using less water, or to develop drought tolerant plant varieties. Another possible threat will come from new pests and diseases. New pests and diseases are always emerging but while changes in climate will themselves put additional stress on crops, they also open up the possibility that agricultural pests may expand their range, or proliferate in the favorable conditions brought about by changes in climate. Coakley et al. (1999) says that climate change could alter stages and rates of development of the pathogen, modify host resistance, and result in changes in the physiology of host-pathogen interactions. On a general scale, the IPCC (2007) reports that at lower latitudes, especially in seasonally dry and tropical regions, crop productivity is projected to decrease for even small local temperature increases (1–2 °C), which would increase the risk of hunger. Lobell et al. (2008) estimates that, the effect of climate change on agriculture being greatest in sub-tropical and tropical areas, southern Africa could lose more than 30 % of its main crop, maize, by 2030, while in South Asia losses of rice, millet and maize could exceed 10 %. The same authors have recently revised these estimates upwards and suggested losses of maize production in Africa of up to 50 % (Lobell et al. 2011).

- **Global warming would also create entirely new climates:** Williams et al. (2007) looked at predicted shifts in climate under different climate change scenarios, and then analysed whether that particular pattern exists anywhere else: novel climates (or new climate not currently seen anywhere on earth) are projected to develop primarily in the tropics and subtropics, whereas disappearing climates are concentrated in tropical mountain regions and the pole ward portions of continents. Under the high-end A2 scenario, 12–39 % and 10–48 % of the Earth's terrestrial surface may respectively experience novel and disappearing climates by 2100 AD. Corresponding projections for the low-end B1 scenario are 4–20 % and 4–20 %. Some land area will be subject to a new climate not seen anywhere within 500 km. This “local” area is greater and reflects also the increased risk to species that cannot disperse further than 500 km (Williams et al. 2007).

More than ever before, the wise use of plant genetic resources will be required.

1.2 How can Agriculture Meet Those Challenges?

1.2.1 *Changes Needed in Agricultural Systems*

In view of this global situation, agricultural systems will need to adapt. They will need to produce more and better quality food under harsher conditions while preserving and protecting the environment. As climate change impacts are predicted to be higher in tropical regions, developing countries' agriculture systems are even most at risk, and they are least able to cope (IAASTD 2009). The genetic diversity contained in crops and the agrobiodiversity contained in farming systems are important elements to consider in a low input and sustainable agricultural system.

1.2.2 *Use of Genetic Diversity and Agrobiodiversity*

Genetic diversity is a source of traits for increased productivity and resistance to biotic and abiotic stresses (Sthapit et al. 2010a). More emphasis has to be put in breeding climate-resilient varieties. Climate-resilient varieties are crop varieties with greater resistance to abiotic stresses such as drought, flooding and extreme temperatures. The genetic diversity of crop wild relatives (CWR)—a wild plant taxon that has an indirect use derived from its relatively close genetic relationship to a crop (Maxted et al. 2006)—and traditional varieties could also be used more extensively as they contain resistance genes for breeding climate-resilient varieties (Hajjar and Hodgkin 2007). Local knowledge to guide crop and variety selection should also be used (Sthapit et al. 2010b, Shanthakumar et al. 2010).

Agrobiodiversity can be used as a tool to minimize risk, especially in marginal areas: more complex ecosystems need to be maintained as they buffer the effects of natural disasters (Holt-Gimenez 2002). Species and genetic diversity is also used to reduce risks of disease outbreaks and mitigate impacts of natural disasters (Zhu et al. 2007, Molina and Molina 2009), long term climate change or unpredictable environmental conditions (Sawadogo et al. 2005, 2006).

1.2.3 *Genetic Resources in Details*

Scientific advances such as high-throughput sequencing, marker-assisted selection and direct manipulation of the genome have allowed breeders to identify traits and incorporate them into improved varieties more efficiently and more rapidly (Varshney and Tuberosa 2007, Tuberosa et al. 2011). The problem is that the genetic resources that are the foundation of all improvements and allow for these efforts in scientific advances are not being managed effectively.

Ex-situ genetic resources collections are currently scattered across more than 1750 genebanks for a total of over 7.4 million accessions worldwide (FAO 2010b). Many of those genebanks are in poor physical condition and continue to be degraded as a result of insufficient and insecure funding. Many of the accessions they conserve are multiple duplicates, which is a waste of precious resources. Moreover, there is little publicly available information about those accessions.

The research Centres of the Consultative Group on International Agricultural Research (CGIAR) actually hold 42 *ex-situ* collections of genetic resources in eleven genebanks, for a total of more than 690 000 accessions (SINGER 2011²). Crop wild relatives, which, as said before, are a precious source of traits for resistance to biotic and abiotic stresses, are poorly represented in genebanks and in any case need also to be conserved in the wild so that they can continue to evolve in response to those stresses. CWR have specific characteristics such as recalcitrant seeds, dormancy, difficult germination, which makes their conservation *ex situ* technically more difficult and complex (Hunter and Heywood 2010). Those wild plants are, at the same time, threatened by climate change. In fact, the predicted rise in global temperatures over the next decades and the consequent changes in rainfall patterns will have a significant impact on the survival of CWR, accelerating the reduction of suitable habitats and increasing the rate of habitat fragmentation, with many predicted to be extinct by 2050 (Bioversity International and UNEP-GEF 2011). Climate change threats to wild potato and wild peanuts populations are being studied based on a model developed by Jarvis et al. (2010), where areas of greatest threats to ecosystems were identified using geographical information system (GIS) data (including deforestation, agriculture, urban expansion, etc).

The Crop Wild Relatives project, funded by the Global Environmental Facility (GEF) of the United Nations Environmental Programme (UNEP) and implemented by Bioversity International in collaboration with the Governments of Armenia, Bolivia, Madagascar, Sri Lanka and Uzbekistan, affirms that there is an urgent need to identify priority species and areas for conservation, to target collecting and to develop integrated conservation strategies to ensure that the rich genetic diversity of crop wild relatives is protected for the benefit of future generations (Bioversity International and UNEP-GEF 2011). The Project was carried out between 2004 and 2011 and sets out to establish a broadly-based partnership to enhance the *in situ* conservation of Crop Wild Relatives in these five countries, and to use the experience of doing so as a platform to create and test tools that would enable others to use similar methods, adding to the global knowledge about CWR and their conservation and use. The countries participating in the project contain some of the world's biodiversity hotspots and most of them had identified wild relatives as a priority for conservation but were unable to do much in the way of conservation because they lacked knowledge and resources.

The project met its goals of establishing working relationships between institutions to promote the development of national and international CWR inventories; prioritizing CWRs, collecting information on their conservation status and use, and

² SINGER. <http://singer.cgiar.org/index.jsp> (last accessed: 22 November 2011).

evaluating their potential for crop improvement: as a result of this effort, species from 36 genera were earmarked for action and more than 310 species were red-listed according to the International Union for Conservation of Nature (IUCN) guidelines; providing effective models for protected areas and species management that can be used in other parts of the world; developing awareness programs for policy makers and other audiences that highlighted the biodiversity of CWRs and their importance for livelihoods and food security; producing tools such as the global CWR portal³ and a manual on *in-situ* conservation⁴. The project substantially expanded the world's knowledge of crop wild relatives, especially in developing countries. It included an ambitious assessment of the distribution, use and threats to these wild species. It organized existing national-level information and made it available on the CWR Global Portal. Some CWRs were assessed for their use in crop improvement for sustainable livelihoods and food security (UNEP/GEF 2011).

1.3 A Global System for the Conservation and Sustainable Use of Plant Genetic Resources

There is an urgent need to address the above-mentioned issues by building an effective global system for the conservation and sustainable utilization of plant genetic resources for food and agriculture.

1.3.1 Today's Situation

As mentioned above, about 7.4 million accessions of crops are today conserved in more than 1750 genebanks over the world (FAO 2010b). Many of those genebanks, however, are underfunded and it is estimated that between 70 to 75 % of the total holdings are multiple duplicates of accessions held in the same or, more frequently, in another genebank. At the same time, other holdings are not safely duplicated (FAO 2010b). Moreover, there is poor information on many accessions and poor access to the available information. Access to that diversity is therefore often problematic for direct use by farmers and for breeders. The global system is actually lacking the strong commitment and a common framework for international collaboration that are necessary to be efficient and to allow effective conservation and use of plant genetic resources, so important for food security and agriculture sustainability.

³ <http://www.cropwildrelatives.org/>.

⁴ Hunter D. and Heywood V. (Eds). 2010. Crop Wild Relatives: A Manual of in situ Conservation. Issues in Agricultural Biodiversity. Earthscan. 440 p.

1.3.2 An Evolving Global System: Some Historical References

Since 40 years, the conservation and use of plant genetic resources has brought into collaboration many organizations and institutions from different countries around the world and has taken several forms, including the establishment of institutional mechanisms and structures, information sharing, the identification of priority activities and the creation of frameworks that can further strengthen cooperation (Halewood and Nnadozie 2008, Hodgkin et al. 2012). The following (Table 1.1) presents, in chronological order, some of the elements which form part of a global system for the conservation and sustainable utilization of plant genetic resources for food and agriculture (PGRFA) today.

1.3.3 Elements of a Global System

The objectives of the global system are the conservation of plant genetic resources useful for food and agriculture, their sustainable utilization and their universal availability. The Food and Agriculture Organization of the United Nations' (FAO) vision of a global system (FAO 1996, 1998) includes three different elements classified as policy, technical and financial elements.

1.3.3.1 The Policy Elements

The policy elements include a legal framework and an overall agreed agenda. The legal framework, the *International Treaty on Plant Genetic Resources for Food and Agriculture* (ITPGRFA or the *Treaty*, for short), is the result of the renegotiation of the 1983 *International Undertaking on Plant Genetic Resources*, which started in 1994. The ITPGRFA was finally adopted by the FAO Conference in 2001 and, unlike the *International Undertaking*, is a legally binding treaty. The ITPGRFA entered into force on 29 June 2004 and covers all plant genetic resources for food and agriculture, recognizes farmers' rights and establishes a Multilateral System of Access and Benefit Sharing (Multilateral System) to facilitate access to plant genetic resources for food and agriculture included in Annex 1 of the ITPGRFA, and to share the benefits derived from their use in a fair and equitable way.

In October 2006, the eleven CGIAR Centres holding *ex situ* collections of plant genetic resources signed agreements with the Governing Body of the ITPGRFA placing the collections they hold under the Treaty. Consequently, the genetic resources they hold are distributed, as of 1st January 2007, using the Treaty's Standard Material Transfer Agreement (SMTA), which was adopted by the Governing Body of the ITPGRFA at its First Session in June 2006. By signing those agreements, the CGIAR Centres commit themselves to support and implement the *Treaty*, and in particular, to work with the international community to build a strong and effective Multilateral System (SGRP 2009).

Table 1.1 Some historical references related to the development of the global system for the conservation and sustainable utilization of plant genetic resources. (Adapted from CGRFA 2011 and from Hodgkin et al. 2012)

Date	Collaborative event
1974	The <i>International Board for Plant Genetic Resources</i> (IBPGR) was established by the CGIAR
1983	The FAO Conference adopts the <i>International Undertaking on Plant Genetic Resources</i> . At the time of its adoption, the <i>International Undertaking</i> , which also lays the foundation for the CGRFA, is the only international instrument specifically dealing with genetic resources for food and agriculture The <i>Commission on Plant Genetic Resources</i> (former CGRFA) is established The development of the <i>Global System on Plant Genetic Resources</i> begins with the establishment of the Commission
1987	Creation of the CGIAR <i>Inter-Centre Working Group on Genetic resources</i> (ICWG-GR)
1989	The CGRFA calls for the development of the <i>International Network of Ex Situ Collections</i> under the Auspices of FAO, in line with the <i>International Undertaking</i> , because of lack of clarity regarding the legal situation of the <i>ex situ</i> collections
1991	The FAO Conference recognizes the sovereign rights of nations over their plant genetic resources
1992	Adoption of the <i>Convention on Biological Diversity</i> (CBD)
1993	The FAO Conference also adopts the <i>International Code of Conduct for Plant Germplasm Collecting and Transfer</i> , developed by FAO and negotiated through the Commission The Commission endorses the <i>Genebank Standards</i> , developed by an expert consultation in 1992, and requests for the preparation of a rolling <i>Global Plan of Action on Plant Genetic Resources for Food and Agriculture</i> (GPA), in order to identify the technical and financial needs for ensuring conservation and promoting sustainable use of plant genetic resources
1994	Establishment of the <i>System-wide Genetic Resources Programme</i> (SGRP) Eleven Centres of the <i>Consultative Group on International Agricultural Research</i> (CGIAR), and subsequently other institutions sign agreements with FAO, placing most of their collections (some 500,000 accessions) under the auspices of FAO. Through these agreements, the Centres agree to hold the designated germplasm “in trust for the benefit of the international community.” The agreements provide an interim solution, until the revision of the <i>International Undertaking</i> has been completed
1995	The FAO Conference broadens the Commission’s mandate to cover all components of biodiversity of relevance to food and agriculture. It renames the Commission the <i>Commission on Genetic Resources for Food and Agriculture</i> (CGRFA)
1996	The first <i>State of the World’s Plant Genetic Resources for Food and Agriculture</i> is presented during the <i>International Technical Conference on Plant Genetic Resources</i> , held in Leipzig, Germany. The Conference welcomes the report as the first comprehensive worldwide assessment of plant genetic resources for food and agriculture. The Conference also adopts the <i>Global Plan of Action for the Conservation and Sustainable Utilization of Plant Genetic Resources for Food and Agriculture</i> , negotiated by the CGRFA, and the <i>Leipzig Declaration</i>
1997	The CGRFA establishes, as “sectoral working groups”, the <i>Intergovernmental Technical Working Group on Animal Genetic Resources for Food and Agriculture</i> and the <i>Intergovernmental Technical Working Group on Plant Genetic Resources for Food and Agriculture</i> to deal with specific matters in their areas of expertise
1998	The first <i>State of the World’s Plant Genetic Resources for Food and Agriculture</i> is published by FAO

Table 1.1 (continued)

Date	Collaborative event
2001	<p>The CBD adopts its <i>Programme of Work on Agricultural Biodiversity</i></p> <p>The FAO Conference adopts the <i>International Treaty on Plant Genetic Resources for Food and Agriculture</i>. This legally binding treaty covers all plant genetic resources for food and agriculture. The Treaty recognizes <i>Farmers' Rights</i> and establishes a <i>Multilateral System</i> to facilitate access to plant genetic resources for food and agriculture, and to share the benefits derived from their use in a fair and equitable way</p>
2003	<p>Establishment of the <i>Global Crop Diversity Trust</i>, which is a resource mobilization mechanism for an endowment fund that will support the <i>ex situ</i> conservation of key crop collections in perpetuity</p>
2004	<p>The <i>International Treaty on Plant Genetic Resources for Food and Agriculture</i> enters into force on 29 June 2004</p>
2006	<p>The First Session of the <i>Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture</i> is held in Madrid, Spain. In accordance with Article 15 of the <i>International Treaty</i>, 11 Centres of the <i>Consultative Group on International Agricultural Research</i> (CGIAR) and other international collections place their <i>ex situ</i> genebank collections under the <i>International Treaty on Plant Genetic Resources for Food and Agriculture</i>. The Article 15 agreements replace the former agreements concluded between the Centres and FAO in 1994</p>
2007	<p>The Second Session of the <i>Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture</i> is held in Rome, Italy</p> <p>The Commission adopts its <i>Multi-Year Programme of Work</i>, a rolling 10-year work plan covering the totality of biodiversity for food and agriculture</p>
2008	<p>The CBD revises its <i>Programme of Work on Agricultural Biodiversity</i></p> <p>Opening of the <i>Svalbard Global Seed Vault</i></p>
2009	<p>The FAO Conference adopts Resolution 18/2009 stressing the special nature of genetic resources for food and agriculture in the context of the negotiations of the <i>International Regime on Access and Benefit-sharing of the Convention on Biological Diversity</i>.</p> <p>The <i>Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture</i> is published</p> <p>The CGRFA adopts the <i>Strategic Plan 2010–2017</i> for the implementation of the <i>Multi-Year Programme of Work</i></p> <p>In view of preparations of the <i>State of the World's Forest Genetic Resources</i>, the Commission establishes the <i>Intergovernmental Technical Working Group on Forest Genetic Resources</i>.</p> <p>The <i>Third Session of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture</i> is held in Tunis, Tunisia</p>
2010	<p>Decision no. X/34 at the <i>CBD Tenth Conference of the Parties</i> (COP-10) drew attention to the importance of work on crop wild relatives and agreed on collaboration with the CGRFA, the ITPGRFA and the FAO on identified activities. At COP-10, parties agreed to adopt the <i>Nagoya Protocol on Access and Benefit Sharing</i></p>
2011	<p>The <i>Fourth Session of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture</i> is held in Bali, Indonesia</p> <p>The FAO Council adopts the <i>Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture</i></p>

The agreed agenda is the *Global Plan of Action on Plant Genetic Resources for Food and Agriculture* (GPA), which was adopted in Leipzig in 1996, since then revised and updated. The Thirteenth Regular Session of the Commission on Genetic Resources for Food and Agriculture (CGRFA) agreed upon the *Second Global Plan of Action for Plant Genetic Resources for Food and Agriculture* in July 2011 (FAO 2011b).

1.3.3.2 The Technical Elements

The technical elements of the global system include global analyses and a global information system. The above mentioned policy elements of the global system are based on technical elements or global analyses such as the *First and Second Reports of the State of the World on Plant Genetic Resources for Food and Agriculture*, which were published in 1998 and 2010 respectively (FAO 1998, 2010b).

The global information system is central to the availability of plant genetic resources. The development of the Internet in the last fifteen years has enabled national and international genebanks to ensure global availability of information about their holdings (Hodgkin et al. 2012). Recent developments aimed at supporting the documentation and exchange of genebank information include the release of GRIN-Global, a genebank management information system with built-in networking features, and GENESYS⁵, a plant genetic resources portal that gives breeders and researchers a single access point to information on about a third of the world's germplasm accessions (around 2.4 million accessions) held in hundreds of genebanks around the world, including those held in the international collections managed by the CGIAR centres, the Germplasm Resources Information Network (GRIN)⁶ of the United States Department of Agriculture (USDA), and the European Internet Search Catalogue (EURISCO⁷) (FAO 2011b). GENESYS is supported by the Secretariat of the ITPGRFA, the Global Crop Diversity Trust (GCDDT) and the CGIAR.

The CGIAR System-wide Information Network for Genetic Resources (SINGER)⁸ has provided invaluable experience and a model to move towards an approach to information management to underpin a global system of plant genetic resources conservation and use. Being developed by Bioversity International on behalf of the CGIAR System-wide Genetic Resources Programme (SGRP), GENESYS takes the purpose and functionality of SINGER to the next level by incorporating data from genebanks outside the CGIAR system and, as well as passport data, including environmental, characterization and evaluation data for each accession (GENESYS⁹). Environmental data from the site of origin enables GIS analysis and the matching of traits to climatic conditions. GENESYS will promote the use of

⁵ <http://www.genesys-pgr.org>.

⁶ <http://www.ars-grin.gov/>.

⁷ <http://eurisco.ecpgr.org/static/index.html>.

⁸ <http://singer.cgiar.org/>.

⁹ <http://www.genesys-pgr.org>.

accessions by allowing users to make queries across all categories of data and place orders online.

The foundation for GENESYS comprises aggregated data from SINGER, EURISCO and GRIN. The evolution of SINGER into GENESYS, looking beyond the CGIAR Centres to the wider community of partners, is in line with and supportive of the larger vision behind the ongoing CGIAR change process. Moreover, by providing an informatics portal that is global in scope, GENESYS is a potential basis for the global information system envisioned by Article 17 of the ITPGRFA.

SINGER dates back to 1994 and is an integrated system to facilitate the management and sharing of genetic resources information relating to the CGIAR collections (eleven genebanks containing more than half a million samples of crop, forage and tree diversity) (SINGER 2011). EURISCO is a web-based catalogue that provides access to information about *ex situ* plant collections maintained in Europe and it receives data from the national European inventories. The EURISCO Catalogue contains passport data on more than 1 million samples of crop diversity from 40 countries. These samples of crop diversity represent more than half of the *ex situ* accessions maintained in Europe and roughly 14% of total worldwide holdings (EURISCO 2011). GRIN provides germplasm information about plants, animals, microbes and invertebrates collections maintained in the United States of America (USA). GRIN provides National Genetic Resources Program (NGRP) personnel and germplasm users continuous access to databases for the maintenance of passport, characterization, evaluation, inventory, and distribution data important for the effective management and utilization of national germplasm collections.

Collaborative activities

To reach the global system objectives of effective worldwide conservation and sustainable utilization of PGRFA, different activities have to be carried out in a collaborative manner at national, regional and international level.

Collaboration on conservation Examples of activities in conservation of PGRFA are the location of the crop diversity and the prioritization of the diversity that needs to be conserved; the collection and analysis of the diversity; the characterization and evaluation of the collected material; the planning of conservation activities (*ex situ* and/or *in situ*) and the effective management and maintenance of the material (Hodgkin et al. 2012). Most of those efforts to conserve PGRFA are undertaken at national level, by organizations and institutions constituting national programmes. National genebanks conserve about 90% of the *ex situ* conserved materials held around the world (FAO 2010b). *In situ* conservation of crop wild relatives and on-farm conservation efforts are undertaken as part of the work of national conservation programmes or directly by farmers and rural communities. In fact, national germplasm collections, once placed under the *Multilateral System of Access and Benefit-sharing* of the ITPGRFA, become a global resource and are as much part of the global system as international collections can be (Hodgkin et al. 2012).

Regional and international collaboration complement and support the work of national programmes by facilitating some international aspects of their work. FAO mentions eighteen regional and sub-regional multicrop plant genetic resources network around the world (such as the European Cooperative Programme for Genetic Resources (ECPGR), the East African Plant Genetic Resources Network (EAP-GRIN), the South Asia Network on Plant Genetic Resources (SANPGR), etc) having a very important role in promoting cooperation, sharing knowledge, information and ideas, exchanging germplasm and for carrying out joint research and other activities (FAO 2010b). Some of those activities include the management of regional *ex situ* collections, such as the Nordic genebank, which is managed by the Nordic Genetic Resources Centre (NordGen); the collections of the Southern African Development Community Plant Genetic Resources Centre (SPGRC); A European Genebank Integrated System (AEGIS)¹⁰, which aims to establish a virtual European Genebank Collection, to be maintained in accordance with agreed quality standards, and to be freely available in accordance with the terms and conditions set out in the *International Treaty on Plant Genetic Resources for Food and Agriculture*.

At the international level, the eleven CGIAR Centres' genebanks, the World Vegetable Centre (AVRDC)¹¹ and the Tropical Agricultural Research and Higher Education Center (CATIE)¹² collections, holding some 700,000 accessions of germplasm, make probably the most substantial direct international contribution to the global system's conservation objectives (Hodgkin et al. 2012). The Svalbard Global Seed Vault (SGSV), which was inaugurated in February 2008, holds some 716,000 seed samples as of December 2011¹³, and acts as a global repository of last resort, or ultimate safety back-up, from seed collections around the globe.

Collaboration on utilization Example of activities related to the availability and sustainable utilization of PGRFA are the selection, breeding and the development of new varieties, the management and dissemination of information about the conserved germplasm, the distribution and exchange of the conserved material, and the maintenance of diverse farming systems. The ITPGRFA and its multilateral system provides the agreed framework of procedures and practices for parties, international organizations and legal and natural persons to make Annex 1 material universally available. Collaboration on use is essentially crop based and undertaken as part of crop networks and activities are often taking place at country level (Hodgkin et al. 2012).

Some example of collaborative initiatives on utilization include the CGIAR Generation Challenge Programme (GCP), the research component on "Enhancing the conservation and use of agricultural biodiversity" (or Agrobiodiversity component, for short) to be included in the CGIAR Research Programme 1.1 on "Integrated Agricultural Production Systems for the Poor and Vulnerable in Dry Areas", the Borlaug

¹⁰ <http://aegis.cgiar.org/>.

¹¹ <http://www.avrdc.org/>.

¹² http://www.catie.ac.cr/magazin_ENG.asp?CodIdioma=ENG.

¹³ <http://www.nordgen.org/sgsv/>.

Global Rust Initiative (BGRI), and the Global Initiative on Late Blight (GILB). Those initiatives are presented below.

The CGIAR Generation Challenge Programme The Generation Challenge Programme of the CGIAR is a global partnership for exploring plant genetic diversity and developing—for resource-poor farmers in harsh environments—crops with improved stress tolerance, with a key focus on improving plants to withstand drought and other stresses such as pests and diseases. The GCP works with more than 200 partners spread across 54 countries. The research activities include the selective characterization of the diversity of the most important crop germplasm for agriculture, including collections stored in gene banks under the custody of the CGIAR as well as country research programmes. Using this diversity, GCP applies genomic tools and interdisciplinary approaches to better understand gene function and gene interactions. This understanding of gene systems across crops helps to identify and tag genes which contribute desired agronomic traits. Selection of favourable alleles or variants of those genes increases the efficiency, speed and scope of plant breeding. GCP also integrates information components and analysis tools into a coherent information gateway and provides support for data storage and analysis¹⁴.

The Agrobiodiversity Component The component on “Enhancing the conservation and use of agricultural biodiversity” is a research programme that was developed by the International Center for Agricultural Research in the Dry Areas (ICARDA), the International Potato Center (CIP), the International Livestock Research Institute (ILRI) and Bioversity International and that will be included in the CGIAR Research Programme (CRP) 1.1 on “Integrated Agricultural Production Systems for the Poor and Vulnerable in Dry Areas”. The development of the Agrobiodiversity component followed the recommendations made by the Genetic Resources Scoping Study report (Qualset et al. 2011), which was commissioned by the Consortium Board of the CGIAR in 2010 to investigate whether genetic resources research and conservation activities were sufficiently incorporated in the new research programmes of the CGIAR, whether there were genetic resources-related cross-cutting issues that had not been addressed or had been duplicated in several CRPs. The study found that a number of cross-cutting research and other issues relevant to the conservation and use of agricultural biodiversity were not adequately covered in current CRPs where the CGIAR should undertake work. The Agrobiodiversity component intends to respond to those identified gaps and proposes to implement work in the areas of *in situ* conservation of crop wild relatives and on farm management of plant genetic resources, which includes a specific sub-theme on facilitating use of landraces, with special attention to neglected and underutilized species (NUS). It also includes collaboration on information and knowledge on plant genetic resources to facilitate access to appropriate information about genetic materials conserved *in situ* to not only scientists and breeders but also to farmers and development practitioners; as well as collaboration on strategies and policies to support the conservation, availability and use of plant genetic resources from local to global levels. The work will be undertaken

¹⁴ <http://www.generationcp.org/>.

in partnership with many different institutions and organizations, more specifically with relevant CGIAR Centres; National Agricultural Research Services (NARS) such as the Indian National Bureau of Plant Genetic Resources (NBPGR-India), the Agricultural Research Service of the USDA (USDA-ARS), The *Empresa Brasileira de Pesquisa Agropecuária* (EMBRAPA), the Chinese Academy of Agricultural Sciences (CAAS); advanced research institutions such as AGROPOLIS, Birmingham University, botanic gardens; global agencies and international organizations working on agricultural biodiversity such as FAO, the Convention on Biological Diversity (CBD) and IUCN¹⁵.

The Borlaug Global Rust Initiative The Borlaug Global Rust Initiative (BGRI), founded by the late Dr. Norman E. Borlaug, replaces the Global Rust Initiative (GRI), which was initially organized by the International Center for Maize and Wheat Improvement (CIMMYT) and ICARDA. The BGRI was established following a recommendation made in 2005 by the Expert Panel on the Stem Rust Outbreak in Eastern Africa in the report “An assessment of race Ug99 in Kenya and Ethiopia and the potential for impact in neighboring regions and beyond” (CIMMYT 2005).

Rusts (*Puccinia* spp.) are present since historical times and many rusts epidemics broke out over the past 150 years, in the near and far east, Europe, and the Americas, causing major famines in Asia and grain losses at a massive scale in North America in 1903 and 1905 and 1950–54. For several decades the problem of wheat stem rust has been resolved through the use of genetic resistance. In Eastern Africa, that resistance has now been overcome by a new physiological race of the disease designated as Ug99. With the long distance travel of rust spores, stem rust reached Kenya in 2001, was established in Ethiopia in 2006, and reached also Sudan and Yemen, reached Iran in 2007 and spread also south to South Africa¹⁶. It is only a matter of time until Ug99 reaches across the Saudi Arabian peninsula and into the Middle East, South Asia, and eventually, East Asia and the Americas. The overarching objective of the BGRI is to reduce wheat vulnerability to stem, yellow, and leaf rusts, support the evolution of a sustainable international system to contain the threat of wheat rusts and continue the enhancements in productivity required to withstand future global threats to wheat¹⁷. BGRI includes partners from around the world. Their website lists a network of 474 professionals working on rusts.

The Global Initiative on Late Blight (GILB) The Global Initiative on Late Blight (GILB)¹⁸ is a worldwide concerted response to potato late blight (*Phytophthora infestans*), the most devastating potato disease that threatens potato crops worldwide. GILB’s primary aim is to improve management of late blight in developing countries. It is a worldwide network of researchers, technology developers and agricultural knowledge agents that serves as a platform to exchange ideas and opinions,

¹⁵ In 2012, it was decided to implement these activities through different CRPs, rather than including them all in CRP 1.1.

¹⁶ <http://wheatrust.cornell.edu/>.

¹⁷ <http://globalrust.org/traction/project/about>.

¹⁸ <https://research.cip.cgiar.org/confluence/display/GILBWEB/Home>.

and facilitates communication and access to information. The research implemented by its partners around the world concern resistance breeding, pathogen studies and disease management. As an example, CIP, one of the CGIAR research centres, is undertaking research on late blight to develop, adapt, and integrate technologies for the management of late blight (LB), the most devastating potato disease in Latin America. Conventional and molecular breeding methods are used to produce advanced breeding populations and clones with durable resistance to LB. Molecular tools are used to characterize the genetic structure of pathogen populations. Additional component technologies are being developed for disease management under the conditions encountered by resource-poor farmers. Integrated disease management methods are being designed and implemented through collaboration with national research systems, governmental and non-governmental extension agencies, and farmers. Agricultural information systems and farmer knowledge are studied with participatory research methodology. In November 2009, CIP coordinated a meeting in Bellagio, Italy, uniting scientists from 21 developed and developing countries to plan a global strategy for combating LB disease. A major output of the meeting was a white paper on the global late blight problem¹⁹.

1.3.3.3 The Financial Elements and Mechanisms

The ITPGRFA adopted a Funding Strategy in 2006 (Resolution 1/2006)²⁰. During its Third Session in 2009, the Governing Body of the ITPGRFA welcomed the *Strategic Plan for the implementation of the Benefit-sharing Fund* of the Funding Strategy and agreed that this plan constitutes a basis for resource mobilization for the *Benefit-sharing Fund* by the Secretariat and the Contracting Parties, with a target of mobilizing US\$ 116 million between 2009 and 2014 (FAO 2011c).

The first call for project proposals under the Treaty Funding Strategy covers the period of 2008–2009: eleven projects received grants. The overall goal of the projects were to help to ensure that local farmers have the crop genetic diversity they need to increase production, withstand changes in climate or pests infestations, and provide their families with food and income which, at the same time, will also ensure that our world has the breadth of crop genetic diversity it needs to face the future.

The Call for Proposals for 2010–2011 granted 18 projects. The call integrated two windows: Strategic Action Plans and Immediate Action plans. Here below are some example of granted projects under the window “Strategic Action Plans” which focused on climate change adaptation through the conservation and sustainable use of plant genetic resources for food and agriculture:

¹⁹ CIP. 2010. Late Blight: Action plan for an effective response to a global threat. White paper prepared by the participants to the Bellagio Late Blight Conference, Bellagio, Italy, 16 –20 Nov, 2009. International Potato Centre (CIP). Available at: <http://cipotato.org/publications/pdf/222222.pdf> (last accessed: 19 December 2011).

²⁰ <http://www.planttreaty.org/content/overview-fs>.

- Community based Biodiversity Management for Climate Change Resilience (Nepal);
- Shared management and use of (agro)biodiversity by indigenous and the traditional communities from the semi-arid region of Minas Gerais State as a strategy for food security and to reduce climate risks (Brazil).

Here below are some example of projects under the window “Immediate Action plans” which focused on assisting farmers to adapt to climate change through the management and conservation of plant genetic resources on-farm and their sustainable use:

- Using local durum wheat and barley diversity to support the adaptation of small-scale farmer systems to a changing climate in Ethiopia;
- On-farm conservation and mining of local durum wheat and barley landraces of Tunisia for biotic and abiotic stresses, enhanced food security and adaptation to climate change;
- Participatory conservation & utilization of rice genetic resources for livelihood and food security (Bhutan) (ITPGRFA 2011a).

The *Strategic Plan for the implementation of the Benefit-sharing Fund* of the Funding Strategy of the Treaty recognizes the Global Crop Diversity Trust, officially established in 2004, as an essential element of the Funding Strategy in relation to the *ex situ* conservation and availability of plant genetic resources for food and agriculture. The GCDT aims to ensure the long-term secure funding for the *ex situ* conservation for the world’s most important crop collections, effectively providing grants which will last forever (GCDT 2011). With an endowment fund of about 120 million USD so far, the Trust was able to provide in-perpetuity funding towards the conservation of collections of 15 major food crops in 2010 for a total of 2,093,970 USD (GCDT 2011).

An effective global system also means to ensure the security and the quality of the conserved material: the Trust has supported the regeneration of threatened collections and their safety duplication in other genebanks to ensure their availability for the future. The regeneration activities comprise 56 projects that target 94,996 accessions in 246 collections of 22 crops held by 86 institutes in 77 countries. The projects are mainly bilateral arrangements with national institutes, but include multilateral partnerships in a few cases, where a crop or regional network can provide coordination. By the end of 2010, over 67 % of the targets had been met, with a total of 63,995 accessions (51,868 seed and 12,127 vegetative accessions) successfully regenerated. Due to the risk of disease transfer, the Trust is also helping partners to safety duplicate *in vitro* some 5,872 accessions from vegetatively propagated crops, of which 3,038 have been completed (approximately 60 %), across 26 different institutes (GCDT 2011).

The CGIAR has a major responsibility in conserving international collections of crops important for food security. The CGIAR Consortium Board and the Global Crop Diversity Trust commissioned a study in 2010, implemented in collaboration with the CGIAR centers that host crop germplasm, on the costs of the *ex situ* conservation of the CGIAR collections in order to design a stable funding mechanism for the preservation, regeneration and distribution of the unique germplasm in these

collections. It is the intention of the GCDT, when its endowment reaches the adequate size, estimated at 460 million USD, to ensure stable long-term funding for the conservation costs of the CGIAR collections. In the meantime, a proposal is being developed to ensure stable funding for the CGIAR conservation activities for the next 5 years (from 2012) whereby the CGIAR Fund would provide about 85 % of the funding in the first year to complement the funding provided by the GCDT. The percentage provided by the CGIAR Fund will then decrease as the endowment of the GCDT grows and can cover a greater proportion of the costs.

The distribution of germplasm material is the ultimate objective of the CGIAR genebanks. Details concerning distributions by Centres of both Annex 1 and non-Annex 1 materials are included in reports that were developed through and on behalf of the System-wide Genetic Resources Programme, and presented to the 2nd, 3rd and 4th sessions of the Governing Body of the ITPGRFA²¹. During the period from 1 January 2007 to 31 December 2009, the CGIAR Centres distributed a total of 1.15 million samples of PGRFA. Approximately 84 % of the samples were sent to developing countries or countries with economies in transition, 9.5 % to developed countries, and 6.5 % to other CGIAR Centres. Eighteen percent were distributed by the Centre genebanks and 82 % by Centre breeding programmes.

From August 2008 until December 2009, The Centres also distributed a total of 5,372 samples of non-Annex 1 PGRFA under the Standard Material Transfer Agreement. Of these, 1,949 samples (36 %) consisted of improved material from the breeding programmes, distributed as germplasm under development and 3,423 samples were distributed by the genebanks. Of the total number of samples transferred, 4,244 (79 %) were sent to developing countries, 791 (14 %) to developed countries and 1 (< 1 %) to countries with economies in transition. Most of the remaining 336 samples (6 %) for which distribution was reported were inter-Centre transfers (ITPGRFA 2011b).

All the above mentioned financial supports are nevertheless still too limited in scope and do not yet constitute a coordinated strategic approach that would otherwise provide the resources needed for the system as a whole (Hodgkin et al. 2012).

1.4 Conclusion

Important steps have been taken in the last thirty years in all three elements contributing to the development of a global plant genetic resources system. A legal framework is now in place at the international level: the International Treaty on Plant Genetic

²¹ CGIAR report to the 2nd Session of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture: <ftp://ftp.fao.org/ag/agp/planttreaty/gb2/gb2i12e.pdf>; CGIAR report to the 3rd Session of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture: <ftp://ftp.fao.org/ag/agp/planttreaty/gb3/gb3i15e.pdf>; and CGIAR report to the 4th Session of the Governing Body of the International Treaty on Plant Genetic Resources for Food and Agriculture: <http://www.itpgrfa.net/International/sites/default/files/gb4i05e.pdf>.

resources for Food and Agriculture. However, there are still a number of important countries that have not ratified the Treaty. In addition, most of the countries that have ratified it have not yet put in place the necessary policies, and in some cases made the necessary changes in their national legislation, to fully implement the treaty and as a consequence, the availability of genetic resources is still restrained.

On the technical side, regional and crop-based networks have been established and collaborative activities have been undertaken in a number of areas. Also, the first steps have been taken in the development of the global information system foreseen in Article 17 of the Treaty with the development of GENESYS. But here also, a lot remains to be done. More national collections have to be linked to GENESYS and the amount of characterization and evaluation data in the system has to be multiplied by several orders of magnitude to fulfill its role. This will require significant investments and strong collaboration among partners.

On the funding side, significant progress has also been made in the last few years. The Treaty has established its benefit sharing fund, the GCDT has reached an endowment of 120 million USD and the CGIAR has made significant steps in securing long-term funding for its collections. However, the benefit sharing fund is still far from its objective and the GCDT requires a significant increase in its endowment before it can play its role fully. But most importantly, investments at the national level need to increase to allow the breeders and farmers to have secure access to the germplasm they need and to build the necessary capacity to make full use of the plant genetic diversity.

References

- Bioversity International and UNEP-GEF (2011) Crop wild relatives global portal. Climate change. <http://www.crowildrelatives.org/cwr/threats.html>. Accessed 24 Nov 2011.
- Bruinsma J (2009) The resource outlook to 2050: by how much do land, water and crop yields need to increase by 2050? Paper presented at the FAO Expert Meeting, 24–26 June 2009, Rome, on How to feed the world in 2050. Food and agriculture organization of the United Nations, Rome. <ftp://ftp.fao.org/docrep/fao/012/ak971e/ak971e00.pdf>. Accessed 10 Nov 2011
- CGRFA (2011) History. Commission on genetic resources for food and agriculture. Food and agriculture organization of the United Nations (FAO). <http://www.fao.org/nr/cgrfa/cgrfa-about/cgrfa-history/en/>. Accessed 25 Nov 2011
- CIMMYT (2005) An assessment of race Ug99 in Kenya and Ethiopia and the potential for impact in neighboring regions and beyond. Report submitted by the expert panel on the stem rust outbreak in Eastern Africa. <http://globalrust.org/db/attachments/about/2/1/Sounding%20the%20Alarm%20on%20Global%20Stem%20Rust.pdf>. Accessed 16 Dec 2011
- CIP (2010) Late Blight: action plan for an effective response to a global threat. White paper prepared by the participants to the Bellagio Late Blight Conference, Bellagio, Italy, 16–20 November 2009. International Potato Centre (CIP). <http://cipotato.org/publications/pdf/222222.pdf>. Accessed 19 Dec 2011
- Coakley SM, Scherm H, Chakraborty S (1999) Climate change and plant disease management. *Annu Rev Phytopathol* 37:399–426
- EURISCO (2011) About EURISCO. http://eurisco.ecpgr.org/about/about_eurisco.html. Accessed 29 Nov 2011

- FAO (1996) The FAO global system for plant genetic resources for food and agriculture. FAO focus. FAO. Available at: <http://www.fao.org/FOCUS/E/96/06/06-e.htm>. Accessed 28 Nov 2011
- FAO (1998) State of the world's plant genetic resources for food and agriculture. Report no. 1. Food and Agriculture Organization, Rome
- FAO (2008) Number of hungry people rises to 963 million: high food prices to blame—economic crisis could compound woes. FAO Media Centre. Rome. <http://www.fao.org/news/story/en/item/8836/>. Accessed 9 Nov 2011
- FAO (2009) The state of food insecurity in the world 2009. FAO. Rome. <ftp://ftp.fao.org/docrep/fao/012/i0876e/i0876e.pdf>. Accessed 9 Nov 2011
- FAO (2010a) The state of food insecurity in the world 2010. FAO. Rome. <http://www.fao.org/docrep/013/i1683e/i1683e.pdf>. Accessed 10 Nov 2011
- FAO (2010b) Second report on the state of the world's plant genetic resources for food and agriculture. Commission on genetic resources for food and agriculture. Food and Agriculture Organization of the United Nations, Rome
- FAO (2011a) The state of the world's land and water resources for food and agriculture (SOLAW)—managing systems at risk. Food and agriculture organization of the United Nations. Rome and Earthscan, London
- FAO (2011b) Report of the commission on genetic resources for food and agriculture. thirteenth regular session. Rome, 18–22 July 2011. <http://www.fao.org/docrep/meeting/023/mc192e.pdf>. Accessed 29 Nov 2011
- FAO (2011c) Fourth session of the governing body of the international treaty on plant genetic resources for food and agriculture, Bali, Indonesia, 14–18 March 2011. IT/GB-4/11/Report. <http://www.planttreaty.org/sites/default/files/gb4re.pdf>. Accessed 1 Dec 2011
- GCDT (2011) Annual report 2010. Global Crop Diversity Trust. Rome
- Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* 156:1–13
- Halewood M, Nnadozie K (2008) Giving priority to the commons: the international treaty on plant genetic resources for food and agriculture. In: Tansey G, Rajotte T (eds) *The future control of food: a guide to international negotiations and rules on intellectual property, Biodiversity and Food Security*. Quaker international affairs programme, international development research centre. Earthscan, London, pp 115–140
- Hodgkin T, Demers N, Frison E (2012) The evolving global system of conservation and use of plant genetic resources for food and agriculture: what is it, and where does the Treaty fit in? In: Halewood M, López Noriega I, Louafi S (eds) *Crop genetic resources as a global commons: challenges in international law and governance*. Issues in agricultural biodiversity. Earthscan, London
- Holt-Gimenez E (2002) Measuring farmers' agroecological resistance after Hurricane Mitch in Nicaragua: a case study in participatory, sustainable land management impact monitoring. *Agr Ecosyst Environ* 93:87–105
- Hunter D, Heywood V (2010) *Crop wild relatives: A manual of in situ conservation*. Issues in agricultural biodiversity. Earthscan, London
- IAASTD (2009) *Agriculture at a cross-roads. Global report*. Island Press, Washington, DC
- IPCC (2007) *Climate change 2007, Synthesis Report. An assessment of the intergovernmental panel on climate change*. IPCC, Geneva
- ITPGRFA (2011a) Treaty's Benefit-sharing fund: list of projects approved. International treaty on plant genetic resources for food and agriculture. FAO. ftp://ftp.fao.org/ag/agn/planttreaty/funding/call2010/BSF2010_Projects_approved_web.pdf. Accessed 1 Dec 2011
- ITPGRFA (2011b) Experience of the international agricultural research centres of the consultative group on international agricultural research with the implementation of the agreements with the governing body, with particular reference to the use of the standard material transfer agreement for Annex 1 and Non Annex 1 crops. IT/GB-4/11/Inf. 5. Fourth session of the governing body, Bali, Indonesia, 14–18 March 2011. International treaty on plant genetic resources for food and agriculture. FAO. Rome. <http://www.planttreaty.org/sites/default/files/gb4i05e.pdf>. Accessed 21 Dec 2011

- Jarvis A, Touval JL, Castro Schmitz M et al (2010) Assessment of threats to ecosystems in South America. *J Nat Conserv* 18:180–188
- Lane A, Jarvis A (2007) Changes in climate will modify the geography of crop suitability: agricultural biodiversity can help with adaptation. *J SAT Agric Res* 4:1–12
- Lobell DB, Burke MB, Tebaldi C et al (2008) Prioritizing climate change adaptation needs for food security in 2030. *Science* 319:607–610
- Lobell DB, Bänziger M, Magorokosho C, Vivek B (2011) Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nat Clim Change* 1:42–45
- Maxted N, Ford-Lloyd BV, Jury SL et al (2006) Towards a definition of a crop wild relative. *Biodivers Conserv* 15:2673–2685
- Micronutrient Initiative (2009) Investing in the Future: a united call to action on vitamin and mineral deficiencies. <http://www.unitedcalltoaction.org>
- Molina AB, Molina IR (2009) The use of genetic diversity in managing diseases of crops: experiences in rice and bananas. Paper presented at the FFTC-PCARRD international seminar on development and adoption of green technology for sustainable agriculture and enhancement of rural entrepreneurship, IIRI, Los Baos, Laguna, Philippines, 28 September–2 October 2009
- Qualset C, Hijmans RJ, McGuire PE et al (2011) CGIAR consortium board-commissioned genetic resources scoping study. Submitted to the CGIAR Consortium Board in February 2011
- Sawadogo M, Ouedraogo J, Belem M et al (2005) Components of the ecosystem as instruments of cultural practices in the *in situ* conservation of agricultural biodiversity. *Plant Genet Resour Newsl* 141:19–25
- Sawadogo M, Balma D, Some L et al (2006) Management of the agrobiodiversity under the clinal variation of rainfall pattern in Burkina Faso: the example of okra drought resistance. In: Jarvis D, Mar I, Sears L (eds) *Enhancing crop genetic diversity to manage abiotic stress*, 23–27 May 2005. Budapest, Hungary, pp 18–24
- SGRP (2009) Standard material transfer agreement. CGIAR system-wide genetic resources programme. <http://www.sgrp.cgiar.org/?q=node/171>. Accessed 29 Nov 2011
- Shanthakumar G, Bhag Mal, Padulosi S, Bala Ravi S (2010) Participatory varietal selection: a case study on small millets in Karnataka. *Indian J Plant Genet Res* 23:117–121
- SINGER (2011) SINGER Website. <http://singer.cgiar.org/>. Accessed 29 Nov 2011
- Sthapit B, Padulosi S, Mal B (2010a) Role of on-farm/in situ conservation and underutilized crops in the wake of climate change. *Indian J Plant Genet Resour* 23:145–156
- Sthapit BR, Silwal S, Gyawali S et al (2010b) Participatory plant breeding as a strategy for supporting the assessment, access, use and benefit of traditional crop genetic diversity in the farmer's production system: overview and the case of the Mansara rice (*Oryza sativa* L.) landrace in Nepal. EUCARPIA 2nd Conference of the “Organic and Low-Input Agriculture” section: breeding for resilience: a strategy for organic and low-input farming systems? Paris, pp 1–3
- Thomas DSG, Twyman C, Osbahr H, Hewitson B (2007) Adaptation to climate change and variability: farmer responses to intra-seasonal precipitation trends in South Africa. *Climatic Change* 83:301–322
- Tuberosa R, Graner A, Varshney RK (2011) Genomics of plant genetic resources: an introduction. *Plant Genet Resour* 9:151–154
- UNEP/GEF (2011) In-situ conservation of crop wild relatives through enhanced information management and field application. 20 projects to showcase 20 historic years of environmental finance. UNEP/GEF celebrating twenty years. <http://www.unep.org/dgef/Portals/43/news/facts/CropsWildFinal.pdf>. Accessed 13 Dec 2011
- Varshney RK, Tuberosa R (2007) Genomics-assisted crop improvement: an overview. In: Varshney RK, Tuberosa R, (Eds) “Genomics-Assisted Crop Improvement (Vol. 1): genomics approaches and platforms”. Springer, Dordrecht, pp 1–12
- Williams JW, Jackson ST, Kutzbach JE (2007) Projected distributions of novel and disappearing climates by 2100 AD. *Proc Natl Acad Sci USA* 104:5738–5742
- Zhu YY, Wang YY, Zhou JH (2007) Crop variety diversification for disease control. In: Jarvis DI, Padoch C, Cooper HD (eds) *Managing biodiversity in agricultural ecosystems*. Bioversity International. Columbia University Press, New York, pp 320–337

Chapter 2

Genomic Approaches and Intellectual Property Protection for Variety Release: A Perspective from the Private Sector

J. Stephen C. Smith, Elizabeth S. Jones, Barry K. Nelson, Debora S. Phillips and Robin A. Wineland

Contents

2.1	Critical Needs to Increase Genetic Gain	28
2.2	Intellectual Property Protection	31
2.2.1	Methods of IPP Used in Plant Breeding	32
2.3	Technical Aspects of Obtaining IPP	33
2.3.1	Concerns About the Use of Molecular Markers to Describe Varieties <i>de novo</i>	35
2.3.2	Concerns About the Use of Phenotypic Characteristics to Describe Varieties <i>de novo</i>	36
2.4	Improving the DUS process: The rationale for Change to the Use of Molecular Characteristics	37
2.4.1	Criteria Required for the Development of Standardized Procedures for DUS	38
2.4.2	Evaluation of SNPs and Development of Standardized Procedures for DUS, EDV, and Variety Identification in Maize	39
2.5	Conclusions	42
	References	43

Abstract Genetic gain is a critical means to improve crop production and will increasingly be relied upon to further improve agricultural productivity in ways that are more sustainable. Partly through the use of molecular markers plant breeders have been able to increase the rate of genetic gain by increasing efficiencies in selection for improved performance of agronomic traits. Greater knowledge of the genetic basis of agronomic traits will help breeders to more efficiently explore and harness plant genetic resources including those that are currently exotic. Efficient processes

J. S. C. Smith (✉) · B. K. Nelson · D. S. Phillips · R. A. Wineland
DuPont Pioneer, 7300 NW 62nd Avenue, P.O. Box 1004, Johnston, Iowa 50131, USA
e-mail: stephen.smith@pioneer.com

E. S. Jones
Syngenta Biotechnology, Inc., 3054 East Cornwallis Road,
Research Triangle Park, Raleigh, North Carolina 27709–2257

to obtain intellectual property protection (IPP) are important to allow the private sector to invest in research and product development. Morphological data are currently the criteria by which varieties are judged to meet the criteria for Plant Variety Protection (PVP); similar data also form an important component of patent filings. Molecular markers that are based upon specific Single Nucleotide Polymorphisms, including those surveyed using whole genome sequence data, now provide the basis for intellectual property (IP) systems that are more efficient, precise, cost effective, better supportive of IP, and with true potential for greater harmonization. We report on how such a transition could be undertaken.

Keywords Genetic gain · Intellectual Property Protection (IPP) · Plant Breeders' Rights (PBR) · Plant Variety Protection (PVP) · Utility patents · Molecular markers · Single Nucleotide Polymorphisms (SNPs) · UPOV · Morphology · Phenotype · Genetic distance · Variety Identification · Distinctness · Uniformity and stability (DUS)

2.1 Critical Needs to Increase Genetic Gain

It is possible to roughly approximate the amount of genetic improvement that farmers in Syria have been able to achieve during 9,000 years that have elapsed since the domestication of wheat (*Triticum* spp.). Araus et al. (2007) report yields of 5.78 t/ha for today's Syrian wheat landraces when grown in irrigated plots and 2.46 t/ha when grown in rain-fed plots; none of the plots received fertilizer. These authors also report historical yields calculated from archaeological specimens of (1.61 t/ha and 1.51 t/ha) (mean 1.56 t/ha). Using these data then annual rates of yield increase due to genetic gain were from $(2.46 - 1.56) = 0.9/9000$ t/ha/yr for rain-fed conditions to $(5.78 - 1.56) = 4.22/9000$ t/ha/yr for well-watered conditions (i.e., 0.1 kg/ha/yr to 0.47 kg/ha/yr for rain-fed and for well-watered sites, respectively).

Achieving a more productive agriculture depends with increasing significance upon the ability to maximize the net positive interactions of genotypic x environmental factors that contribute to harvestable yield. Average wheat yields in Great Britain today are 8.5 t/ha and up to 10–14 t/ha. These yield levels far eclipse those that had been achieved in previous centuries and millennia in Great Britain (Fig. 2.1a). It was not until AD 1750 that wheat yields in Great Britain reached levels that had already been achieved nearly 8000 years earlier in the middle Euphrates area. Wheat yields in Great Britain rose at a faster rate over the next two centuries predominantly as a result of changes in crop management practices. In contrast, additional capabilities to raise yields by increasing the rate of genetic gain beyond that which could be accounted for by mass selection only became possible after the rediscovery of Mendel's laws of inheritance at the beginning of the 20th century. Garton Brothers developed a succession of new wheat varieties adapted for Great Britain starting with "White Monarch" in 1899. A significant increase in the rate of yield increase

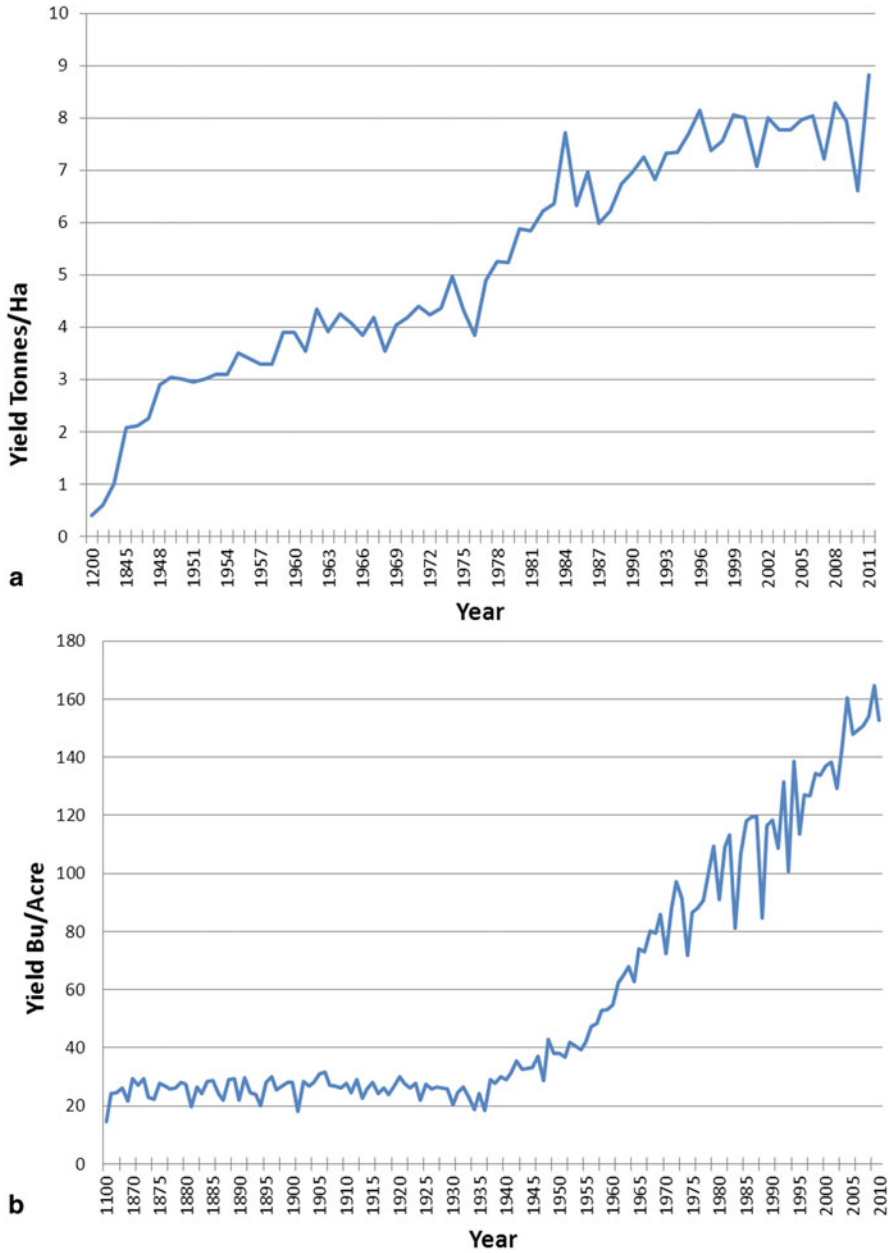


Fig. 2.1 a UK wheat yields (tonnes/ha) from AD 1200–2011 (Austin and Arnold 1989; Ogilvie and Farmer 1997) **b** US maize yields (bushels/acre) from AD 1100–2010. (Smith 1989; FAOSTAT 2011)

occurred during the 1960s due to continued changes in crop management practices coupled with more effective plant breeding. It is estimated that since 1982, 88 % of the gains in cereal yields of UK winter wheat are now due purely to genetic change brought by plant breeding and that “national yields could well be in decline in the absence of any variety improvement.” (Mackay et al. 2011). Following the completion of the UK 2012 wheat harvest yields, a farm survey estimated to be 14 % below the five year moving average due to the wettest growing season in a century (NFU 2012); consequently, wheat and food prices are expected to rise (Malik 2012). UK wheat yields in 2012 were equivalent to lower mean yields last seen as a result of good harvests during the late 1980s (NFU 2012). These data help demonstrate advances in yields that farmers and societies have grown accustomed to, but often without acknowledgement or any appreciation of their causalities, while once again providing a reminder of the confounding effects of climate.

As an extreme example, consider instead that the rate of gain in wheat yields in Great Britain had remained the range of 0.1–0.47 kg/ha/yr, a rate achieved in Syria during the millennia following domestication and without the participation of plant breeders. With those rates of genetic gain it would require from 8,723 to 41,000 years to reach just half the present day wheat yields in Great Britain (c. 8.5 t/ha) from their level in AD 1220 (0.3 t/ha) (assuming the other half of yield improvement to then equal today’s yields would come from improved crop and soil management practices). Under these circumstances inhabitants of Great Britain could anticipate today’s wheat yields would not be reached until the year AD 9943 at the earliest; and by AD 42220 at the latest (and assuming good weather)!

A similar story can be told for the history of maize agriculture in the United States. Maize was initially domesticated near Oaxaca, Mexico, some 8,000 years ago. US maize yields (Fig. 2.1b) remained stagnant from the Civil War until the advent of hybrids in the 1930s. Increases in US maize production prior to the 1930s were entirely due to the taking of more land into agriculture. In contrast, yield changes during the last eight decades have been due to changes in crop management (improved weed and pest control, fertilizer, mechanization, and increased planting density) with the selection of genotypes that are better adapted to yield in those changed management conditions (Castleberry et al. 1984; Russell 1984; Duvick 2005). These examples are representative of the broader global picture. Calderini and Slafer (1998) found that yields achieved soon after the beginning of agriculture were very similar to global average yields that were being attained, some 8,000–10,000 years later at the beginning of the 20th century, prior to the large-scale initiation of plant breeding. These data indicate that global demand for cereals had, at least up to nearly one century ago, largely been met by taking more land into cultivation. To continue such an approach is not sustainable (Borlaug and Dowsell 2005). The need to develop new crop varieties optimally adapted to a productive agricultural environment can help reduce pressures on the environment and maintain biodiversity. Regardless of whether one espouses the “land sparing” (minimizing demand for farmland by increasing yield) or “land sharing” (wildlife friendly farming boosting densities of wild populations on farmland, which may decrease agricultural yields) (Green et al. 2005) strategies for maintaining or increasing biodiversity, it is imperative that crop varieties are best adapted by virtue of their genetic potential to the habitats in which they are grown

so as to maximize production in those environments (Green et al. 2005; Godfray 2011; Phalan et al. 2011). The same argument for establishing the most appropriate fit of crop genotype with agricultural environment applies whether farming is conducted organically or with the aid of chemical pesticides and fertilizers. Indeed, it may well be that under organic or low input farming conditions, development of the best adapted crop genotype is especially critical as, in those circumstances, there will likely need to be yet greater reliance on the use of genetic inputs to help counter pressures from pests, diseases, and weeds.

Sustainable increases in agricultural productivity while safeguarding biodiversity will require diverse approaches such as biotechnology, and certain elements used in “organic” farming (Lal 2001; Marlander et al. 2003; Ammann 2008, 2009; Brookes and Barfoot 2008; Rudel et al. 2009; Ronald 2011; Raven 2010; Bennett et al. 2011; Godfray 2011; Foley et al. 2011; Phalan et al. 2011). The Royal Society (2009) has called for the “sustainable intensification of global agriculture in which yields are increased without the adverse environmental impact and without the cultivation of more land.” DEFRA (2009) estimate wheat yields in Great Britain for 2025 and 2050 of 11.4 t/ha and 13.0 t/ha, respectively. “Protecting biodiversity and ensuring food security are part of a single agenda” (Godfray 2011). And while the further intensification of agriculture alone may not ensure environmental sustainability, “it is an essential step in the process because crop and pasture lands comprise about one-third of Earth’s ice-free surface” (Rudel et al. 2009).

Future capabilities to further improve agricultural production will continue to be dependent upon the more effective use of genetic resources to allow farmers to increase productivity in the face of changes in climates, methods of crop husbandry, and to resist ever-evolving strains of pests and pathogens (Hoisington et al. 1999; Crookston 2006; Qin et al. 2006; Warburton et al. 2006; Glaszmann et al. 2010; Peng et al. 2011). Not only will farmers be asked to produce more bountiful harvests under these circumstances, they will also be expected to contribute to improved nutrition, new products for industry including biofuels, and to contribute as stewards of the environment (Foley et al. 2011). Faced with these demands, farmers will increasingly look to a continual supply of new and better adapted varieties. As other avenues for improving crop production plateau (e.g., weed and pest control) or genetic means are found to achieve the same ends, future prospects for continuing to improve productivity will increasingly rely upon the breeding of better adapted genotypes (Mackay et al. 2011).

2.2 Intellectual Property Protection

It is critical that both public and private resources be optimally invested into research and product development required for plant breeding. A strong and effective public sector is vital to provide a global foundation of more basic research, education and for the conduct of plant breeding, most especially in circumstances where the private sector is absent or inadequate to meet farmers’ needs by region and/or by crop. Many public sector institutions seek IPP. The Bayh-Dole Act of 1980 allows US universities

to obtain ownership of an invention in preference to the US government. Universities in many countries have established dedicated groups to facilitate obtaining IP and licensing including in the field of plant breeding and agricultural biotechnology. The John Innes Centre, a research centre with charity status in the UK, states that “our research innovations will often require substantial further investment to reach applications in the market place and that IP protection has an important role in creating favorable conditions for the uptake and use of such research findings” (JIC 2012). The Brazilian Agricultural Research Corporation (EMBRAPA) which is very largely publicly funded uses IPP as do many other publicly funded research centers as a critical element to help them to achieve their goals to increase agricultural productivity (Cohen, 2000). In rare cases, centres of the CGIAR can utilize IP when it can help achieve goals of increased public dissemination of improved varieties or technologies (SGRP 2010). The commercial basis of privately funded organizations makes the protection of IP mandatory; no commercially funded organisation can be sustainable if it provides the fruits of its research free to its competitors without at least first having had the chance to recoup its investments. Our goal here is to address the technical aspects of describing and identifying varieties to help achieve an IP environment that can optimally attract private investments into commercially funded plant breeding with the goals to contribute to the increase of genetic gain and social welfare (Hayes et al. 2009).

2.2.1 Methods of IPP Used in Plant Breeding

There are at least four ways by which plant breeders can obtain IPP. These are (1) contracts, (2) trade secrets, (3) PVP or Plant Breeders’ Rights (PBR), and (4) Utility Patents. In addition, the United States provides PVP-type protection for varieties of asexually reproducing non-tuberosus species (The US Plant Patent Act). Trade secrets can help provide protection for parent lines of hybrids. Inbreeding depression, which occurs as a result of pollination in fields of F1 hybrids, contributes to encouraging farmers to purchase new hybrid seed annually. Contracts can include bag-tag “shrink-wrap” type protection including use in closed-loop systems where growers contract to return harvested seed to the owner of the variety. PVP is a *sui generis* form of protection prescribed by diplomatic conferences of the L’Union internationale pour la protection des obtentions végétales (International Union for the Protection of New Varieties of Plants (UPOV) indexUPOV). Under the auspices of the 1995 Trade-Related Aspects of Intellectual Property Rights (TRIPS) within the World Trade Organisation, countries may exclude plants and animals from patentability. However, any country that does exclude plant varieties from patent protection is obliged to provide an effective *sui generis* system of plant protection. Patent laws are country specific, for example France and Germany provide exemptions to allow further breeding including commercialization of the non-patented germplasm background in circumstances where specific traits are patented whereas US patent law has no such exemption.

Table 2.1 Comparison of plant breeders rights and patent systems. (After Krattiger 2004)

Criteria	UPOV 1978	UPOV 1991	Utility patents	U.S. plant patent act
Protects	Varieties of listed species	Varieties of all species	Plant genotype not normally found in nature	Asexually reproduced nontuberous plants
Requires	Novelty Distinctness Uniformity stability	Novelty Distinctness Uniformity Stability	Novelty Utility Nonobviousness Enablement	Novelty Distinctness Stability
Disclosure	Full morphological description	Full morphological description	Enabling disclosure Best mode disclosure deposit of novel material	As complete as possible Photographs and drawings
Claims			Refer to specific patents	Single varietal claim
Exemption	Farmer and breeder exemption	Farmer and breeder exemption	Some country patent laws include exemptions	None
Rights	Prevents others from producing for commercial purpose	As UPOV 1978 plus prevents import and export and extension to essentially derived varieties	Prevent others from making, using or selling claimed invention	Prevent others from asexually reproducing, selling, or using

UPOV, Union internationale pour la protection des obtentions végétales (International Union for the Protection of New Varieties of Plants) (<http://www.upov.int/>)

Reviews of IP methods are provided by Williams and Weber 1989; Fernandez-Cornejo 2004; Krattiger 2004; Le Buanec 2004; and by CAMBIA (undated) available at <http://www.patentlens.net/daisy/patentlens/1234.html>. An outline comparison of PVP and patent systems (after Krattiger 2004) is presented in Table 2.1.

The International Seed Federation recently completed a revision of its position on IP (ISF 2012).

2.3 Technical Aspects of Obtaining IPP

Practical enforcement of IPP requires that individual varieties have a legally enforceable grant of protection. In order to meet this requirement, varieties must be characterized or described and subsequently be reliably identifiable from other varieties, including from among those that are closely related. It is very highly desirable that variety identification is free from interference by environmental factors that can affect phenotype, possible at any point in the life-cycle of the plant, be fast (minutes to hours) and relatively inexpensive. In practical terms, these demands can only be met through the use of molecular marker data.

Particularly during the last 3–5 years, there have been huge improvements in every conceivable aspect of molecular marker profiling; the genome can be assayed in very high detail (thousands of SNPs), very quickly (hours), with extremely high fidelity, and with direct output to databases linked to agronomic field data and pedigrees (Mackay TFC 2009; Lai et al. 2010; Yan et al. 2011). Whole genome sequence scans of inbred lines and varieties may soon be cost effective and the norm. These methods are increasingly becoming an integral part of plant breeding to help better understand and thus, to manage the genetic control and expression of important agronomic traits. Just as modern sequencing technologies are being used to help characterise the genetic basis of crop germplasm, one might also reasonably suppose that a similar source of data would be used to identify and characterise new varieties for registration, certification, evaluation of varietal purity, and the granting of IPP.

It must therefore seem unimaginably arcane and probably completely nonsensical to most readers, when they discover that plant varieties are still primarily evaluated for eligibility for varietal status and awarding of IPP upon their morphological appearance. For example, a newly developed genotype will be ineligible as a new variety (and thus not be eligible for protection under PVP and rejected from use in agriculture where countries require variety registration) if it cannot be shown to be morphologically distinct from all previously known varieties of that species. Most all of the morphological characteristics that are used to determine distinctness were purposely chosen so as not to be associated with agronomically important features of a cultivated variety.

Molecular marker data (isozymes and protein storage proteins) were first used in the plant breeding industry to characterise inbred lines, hybrids and varieties, to measure varietal purity and to test pedigrees in the early 1980s. Then, as a result of rapid, and enduring series of developments in molecular marker technologies UPOV established a special working group (Biochemical and Molecular Techniques) to provide an ongoing review of the capabilities of molecular marker systems in respect of their potential usage to support the goals of UPOV. Areas where UPOV currently states that marker data can be employed are: (1) variety identification or validation, but only after making initial description; (2) as a perfect surrogate for an existing morphological or resistance characteristic, (3) to help manage reference collections more efficiently, and (4) as a means to help determine genetic conformity in respect of determining whether a variety is essentially derived from an initial variety.

To date, neither UPOV nor seed associations have accepted that marker data alone can be used to help characterise or to describe a new variety *de novo*. Instead, determination that a genotype meets the criteria established by UPOV to be afforded the status of a new variety still depends upon an examination of numerous morphological characteristics in regard to the criteria of Distinctness, Uniformity, and Stability (DUS). There is widespread acceptance that molecular marker data can be used to identify varieties once a *de novo* description has already been made (MMEDV 1999; ESA 2011; Heckenberger et al. 2005a, b, c; ISF 2004a, b, 2005, 2006, 2007a, b, 2008, 2009, 2012; Rodrigues et al. 2008; Kahler et al. 2010).

The very rapidity of marker system development during the past 30 years has itself confounded the ability of authorities to approve an internationally agreed standardized system. Moreover, each marker system has had some drawbacks which have

undermined the desire to advocate, at least until now, for a change from the use of morphological characteristics to the use of molecular data. For example, Restriction Fragment Length Polymorphisms (RFLPs) first made a hundred or more molecular markers available per species with the ability to survey markers at known map positions collectively sampling each chromosome arm. However, RFLPS were heavily resource demanding and relatively slow to use. Simple Sequence Repeats (SSRs) provided a significant step forward in throughput with increased reliability of scoring. However, SSRs scored on one platform could not always readily be aligned with those scored on a different platform and changes in the design of PCR primers had to be very carefully monitored. More recently, Single Nucleotide Polymorphisms (SNPs) have become the marker system of choice due to their compatibility with ultra-high throughput information, chemistry and robotic laboratory systems and to have access to tens of thousands of markers collectively providing a very high power of discrimination with very high repeatability. Another huge advantage of SNPs is that, because they are sequence based, the same designated loci can still be scored even as DNA platform technologies continue to change and including when whole genome sequence data are collected. SNPs therefore provide a culmination of the most efficient, cost effective, discriminatory and enduring marker data that can be used to characterise cultivated varieties of crop species.

Factors such as methylation, which are beyond the realm of characterisation or prediction of simply inherited Mendelian traits, including molecular markers, play as yet an imperfectly understood but important role in the expression of phenotype (Martienssen and Colot 2001). In addition, there is huge underlying complexity of interactive regulatory networks which condition phenotype (Kaufmann et al. 2010). These findings might suggest that an approach using molecular markers to characterize and to distinguish among plant varieties both qualitatively and quantitatively is naïve, if not redundant. We would respond that whether the goals are to evaluate DUS or to identify or compare varieties then the source of data should be selected that most effectively and efficiently allows the relevant criteria to be determined. At least in our experience, it is clear that the continued use of a set of morphological attributes is neither optimal nor practical to most effectively help provide for and support IP for varieties of major field crop species.

2.3.1 Concerns About the Use of Molecular Markers to Describe Varieties de novo

Primary concerns about the use of molecular marker data as evidence upon which to determine varietal status *de novo* relate to (1) the ability to find numerous additional and potentially additional discriminatory data due to even a very low level of residual heterogeneity (and so potentially undermine the protection vested in the initially declared variety); (2) additional costs that would be born by breeders having to run marker assays during the breeding process to also assure that varieties are uniform, and thus stable, for the same characteristics that are used to establish distinctness; (3)

concerns that an unduly high level of uniformity would be required; and (4) the need to establish standardized protocols and to demonstrate that laboratories around the globe can generate data of sufficient quality and consistency. Some have additional concerns which relate to the legal definition of how a variety is described “by the expression of its characteristics” and who consequently argue that since molecular marker data are not expressed, at least in respect of being the result of transcription of the genome, then they should be excluded from eligibility for this usage.

2.3.2 Concerns About the Use of Phenotypic Characteristics to Describe Varieties de novo

Concerns about the continued use of phenotypic characteristics stem, in very large part, from the large contribution that genotype \times environmental interaction and experimental error play in undermining the precision, and thus the discriminative power, of descriptions based upon morphological characteristics. Morphological data were selected as the characteristics to provide the basis for determining varietal status because prior to the mid 1970s no other means had been conceived or developed for any crop species. UPOV understood the problems that genotype by environmental ($G \times E$) interaction, a feature inherently associated with morphological expression, would contribute by undermining the precision of varietal descriptions and thus making harmonization a particularly difficult proposition. UPOV attempted to mitigate these problems by categorizing morphological characteristics according to the presumed level of complexity of their genetic control; characteristics believed to be under simpler genetic control would be less likely to be influenced by interaction with the environment. However, more recent studies of the genetic control for many of these characteristics (that are only now possible due to the availability of numerous molecular markers) indicate that the genetic control of many, if not most, of these characteristics is quite complex ensuring that their expression is quantitative in nature (Sourdille et al. 1996; Austin et al. 2001; Bredemeijer et al. 2002; Mickelson et al. 2002; Enoki et al. 2006; Li et al. 2007; Tian et al. 2011). For example, Tian et al. (2011) have shown that in maize 30–36 quantitative trait loci (QTLs) are associated with just three leaf characteristics (upper leaf angle, leaf length, and leaf width). And, for example, even when identical protocols have been used to collect morphological description data for the same genotypes in different locations, then even those data can be quite dissimilar (Jones et al. 2003; Hof and Reid 2008).

Due in large part to the effects of $G \times E$, significant dedicated outlays of personnel and field resources are required to obtain morphological data that are statistically meaningful for DUS purposes. These issues become exasperated as the numbers of known varieties annually increases, as is the case for most major field crop species, thereby rendering it practically impossible to directly compare a prospective new variety with all previously declared and known varieties. Consequently, additional strategic approaches have had to be developed in an attempt to fulfill the UPOV requirement to compare new potential varieties with varieties of common knowledge in that species. These approaches seek to use molecular marker data or a combination

of marker and morphological data to exclude from the need for further comparisons varieties that are super-distinct from a candidate variety. As a result, a short-list of existing varieties that would then need to be more closely examined to each candidate variety using morphological characteristics is created. Such a pre-screening exercise can help significantly reduce the amount of field testing that is required in a second and final year of field evaluations. It is also a requirement that a new candidate variety is compared to the most similar variety and the distinguishing characteristics identified. Countries or regions that utilize a centralized approach to generating and comparing morphological data can make such comparisons using the database of all previously known varieties of that species available to them. However, in countries or regions that require individual breeders to be the source of morphological data (e.g., the U.S.), then breeders can only report the closest variety comparison in respect of the varieties for which morphological data are present in their own database. It is not realistic to expect that every breeding company should create and maintain a morphological database of all varieties of common knowledge. Indeed, with regard to inbred lines of a hybrid crop that are maintained as trade secrets then only the inbreds that are publicly known or which are proprietary to that individual breeder will be available to that breeder. Consequently, there is a lack of consistency in how the most similar comparison variety is selected. In addition, the collection of data by individual applicants can lead to a variety of specific protocols being developed for the collection of the required morphological traits. In addition, large G×E effects associated with the expression of morphological characteristics serve to inflate variance, which then diminishes the ability to discriminate between varieties. Finally, there are significant differences, for historical reasons, between the manner in which morphological data are collected, databased and compared across countries, most significantly with regard to the U.S. compared to other UPOV member countries. The U.S. requires actual metrical data with standard errors. In contrast, UPOV requires data be translated or normalized to scores. In our experience, while both methods are effective at evaluating distinctness, it is impossible to meaningfully merge the data from different systems of measurement, even after employing a variety of approaches to adjust or normalize the data so that “it speaks the same language” as the other system (Law et al. 2011a, b, c).

2.4 Improving the DUS process: The rationale for Change to the Use of Molecular Characteristics

Some reference collections of many crop varieties, especially with regard to the major field crops, have already become very large therefore making it practically impossible to satisfy the requirement that new varietal candidates should be compared to all other varieties of common knowledge in that species, at least using morphological data. It is not surprising, therefore, that there is growing interest in the use of molecular markers to reduce workloads and possibly also the costs for PVP offices to manage reference collections. In that context, it would be logical to discuss whether the use of such a tool could reasonably be extended to all aspects of

DUS testing. The use of molecular markers has the potential to significantly reduce error that is effectively caused by genotype \times environment interaction and which effectively reduces the precision of a description that can be realized using morphological data. Marker data can also provide a more standardized approach to the definition of distinctness because the genetic control and map positions of marker loci can be understood. Marker data also allow implementation of the requirement that new putative varieties should be compared to all publicly known varieties of the species. And, as already noted, there are considerable differences in the technical implementation of criteria to measure, record, and compare morphological descriptions. Selecting a standard set of molecular markers on a crop by crop basis has the potential to achieve a significantly higher level of harmonization and a more unified approach among national authorities; an aspiration that UPOV ranks highly.

2.4.1 Criteria Required for the Development of Standardized Procedures for DUS

For variety testing on the basis of D, U and S for plant variety rights or for national listing it would seem reasonable that the process could be carried out entirely with data provided by molecular markers, provided that the following basic tenets are met:

- UPOV wide consensus on the use of DNA-based markers in the DUS examination process in order to obtain international acceptance of the DUS examination reports. UPOV members would need to agree on a harmonized approach before implementing the use of DNA-based markers in the different national DUS examination processes. A transitional period may be required.
- There can be no risk of decreasing the minimum distance, necessary for the declaration of distinctness. A standard set of markers would need to be described with thresholds for distinctness such that it would be impossible, in the extreme case, to declare Distinctness on the basis of only one base pair difference in the whole genome.
- Implementation of a marker based system for DUS assessment might not necessarily remove the requirement for the breeder to achieve Uniformity and Stability of morphological characteristics that are important to the farmer, grower or seed certification authorities. At the same time, the introduction of the use of molecular markers should not generate any additional practical constraints for the breeders, e.g., in the field of variety description.
- Implementation of a marker based system for DUS assessment would require a crop by crop approach.

Such implementation would require:

- The use of a set of markers able to provide the highest discrimination capacity (polymorphism information) and genome coverage. It will be necessary to evenly sample the genome.

- The DNA-based markers, the methods to produce and record the data and carry out statistical computations need to be publicly available.
- When used to determine Distinctness, then either the same full set or a subset of these DNA-based markers also need to be used for assessing Uniformity and Stability.
- Evidence shows that there is intra-variety variation for DNA-based markers within existing varieties that have been declared sufficiently uniform and stable. Therefore tolerance levels for marker uniformity and stability would need to be established on a crop by crop basis.
- There can be no significant increase of the application or examination costs due to such implementation.

2.4.2 Evaluation of SNPs and Development of Standardized Procedures for DUS, EDV, and Variety Identification in Maize

Many publications show the advantages afforded by the use of SNPs; high-throughput, high map resolution, and high repeatability (very low error rates) (Tenaillon et al. 2001; Bhatramakki et al. 2002; Ching et al. 2002; Vroh Bi et al. 2006; Jones et al. 2007). Jones et al. (2007) reported data repeatability of, from 98.1 to 99.3 % depending upon platform technology. And while SNPs show, on average, less polymorphism on an individual locus basis due to their bi-allelic nature (compared to SSR loci which are usually polyallelic in populations), this potential limitation can be countered either by (i) selecting SNPs that have relatively high discrimination ability (a high Polymorphic Index Content or PIC) or by (ii) assembling individual linked or adjacent SNP loci into haplotypes and then reporting those individual haplotypes out as individual alleles of a polyallelic system. Some have reported that from 7–11 times the number of SNP markers are required to provide an equivalent degree of discrimination as SSRs (Laval et al. 2002; Yu et al. 2009; Van Inghelandt et al. 2010). In contrast, Nelson et al. (2011) found that, “at least when the issue is to examine genetic similarities among moderately or closely related germplasm, then the appropriate number of SNP loci need may be in the range of 300–400, . . . provided they are selected on the basis of maintaining relatively high H_e (PIC) as well as even genome coverage”.

The American Seed Trade Association (ASTA) and French Maize Breeders Association (UFS) have jointly embarked on a project using an Illumina 56,000 SNP chip to profile a set of maize inbred lines including those of historic and current importance. The results of this study will provide a list of thousands of publicly available SNP loci from which can be selected a set to measure genetic similarities (or distances) between pairs of inbred lines for the purpose of resolving questions in regard to status as an Essentially Derived Variety. Such a set of SNPs could also be used for determining Distinctness. A much smaller subset of SNP loci could also be selected that would be used to evaluate uniformity and stability. The results and

conclusions of this research will be presented to the International Seed Federation (ISF) and consequently could become promulgated by the ISF as recommended procedures. Similar strategic approaches to select SNP sets and subsets could be used for maize and for other crops in various regions of the world. It will be important to select sets and subsets of SNPs based upon their proven ability to discriminate among germplasm that is relevant for each region and thus to avoid ascertainment bias in the selection of the SNPs.

It will be important to establish a minimum level of % SNP profile similarity as a threshold for Distinctness. Such a threshold must be based upon use of a set of SNPs that are identified as meeting the criteria of (1) collectively providing for fairly even genome coverage and (2) having a proven ability to discriminate among varieties, even among those that are very closely related by pedigree. Otherwise, it could always be potentially possible to find at least one (of several million SNPs) that would have a different base, or be polymorphic for the presence or absence of several SNPs in one or more genetic regions (Fu and Dooner 2002; Wang and Dooner 2006). One could consider using two approaches, or an amalgam of the two: First, given repeatability estimates of from 98.1–99.3 % that have been reported to date for the use of SNP technology in maize, then a value of 98–99 % similarity could be an appropriate minimum boundary for determining distinctness. Second, it is also possible to examine SNP similarities among pairs of inbred lines or varieties that have already been declared as distinct varieties on the basis of morphological comparisons. Such comparisons provide an opportunity to declare a distinctness threshold using SNPs that is calibrated using existing varieties and according to previously agreed morphologically based standards used to declare distinctness; i.e. to calibrate a SNP threshold in relation to customary standards using morphological data. SNP data presented by Nelson et al. (2008) for a set of U.S. maize inbred lines for which PVP certificates have now expired (and including several important widely-used publicly bred lines) provide some interesting comparisons. All of the inbreds reported by Nelson et al. (2008) have been declared distinct on the basis of morphological comparisons conducted by the US PVP Office. Variation was observed among different sources of the same inbred line ranging from 96.3 % (for Mo17) to 99.7 % similar (for three replicate Pioneer bred inbred lines). The most similar pair of inbred lines was for B73 compared to the inbred F42 (approximately 99 % similar). Several other pairs of inbred lines also had very similar SNP profiles: B73–DJ7 (96.6 %), PH207–Q381 (approx. 98 %), Mo17–Seagull–17 (approximately 97 %), LH51–Mo17 (92.9 %), B73–NKW8304 88.8 %). These data raise cautionary issues concerning the prospect of introducing marker data into the PVP system that will need to be taken into consideration. First, some of these pairs of different inbreds are so similar by marker data that the equivalent marker threshold for distinctness would have to be impractically high and overlap with typical levels of laboratory error. In these cases a re-examination of the criteria and evidence used to determine distinctness may be warranted. Second, marker-based distances between different seed sources of Mo17 are so high that, if they are to be considered in the development of uniformity thresholds using SNPs then this may lead to an unwarranted decrease in the standards used to declare distinctness. The finding of high levels of heterogeneity

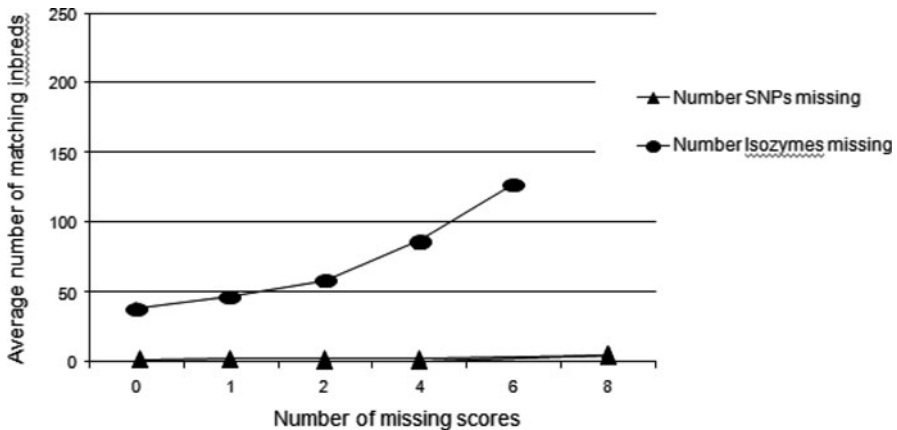


Fig. 2.2 The number of inbreds from a reference set of 438 which have 15/15 matching isozyme profiles or 16/16 matching SNP profiles in relation to the number of missing isozyme or SNP scores. SNPs maintain a much higher power of discrimination even in the face of 50 % missing data

among different seed lots of older inbred lines comes as no surprise because most of these lines were not thoroughly selfed to uniformity before public release. Inbred lines of maize are currently bred to higher standards of uniformity.

It is an important principal that either the same or a subset of characteristics, which are used as the basis for determining distinctness, are also used to assess uniformity, and thus stability. Otherwise, it would be possible to select inbred lines or varieties that meet the threshold of distinctness solely by purification of a heterogeneous seed lot. It is well known that seed lots of inbred lines that have not been selected for uniformity using marker data can still be heterogeneous (Mauria et al. 2000; Mauria et al. 2002; Nelson et al. 2008). Consequently, it will be necessary to develop a panel of SNP loci which can be used cost-effectively during the breeding process, and subsequently also in determining genetic purity of inbred and hybrid seed lot. It is increasingly routine breeding practice in maize to use double-haploids for inbred line development, consequently an increasing proportion of inbred lines and varieties will be completely homozygous and will remain so unless there is contamination or physical mixing during subsequent generations of seed increase or hybrid seed production.

We have found that providing SNP loci are judiciously selected, it is possible to identify a relatively small set of SNP loci that collectively have a very high power of discrimination among maize inbred lines, and which therefore could be used to measure uniformity and stability (genetic purity) e.g., as few as 16 SNPs can discriminate among > 400 Pioneer proprietary inbreds. A set of 16 SNPs can give 16 times the level of discrimination compared with the standard set of 15 isozymes that have historically been used at Pioneer to assess uniformity, and SNPs are far more robust in the face of missing or erroneous data (Fig. 2.2). In addition, where inbreds cannot be discerned using sub-sets of the 16 SNP markers, then those inbreds are likely to be highly related. In contrast, some of the inbreds that could not be distinguished by isozyme data were unrelated by pedigree. Consequently, a small

panel of SNP markers can therefore be cost-effective for routine use during the breeding process, and can provide, not only an initial test of the level of uniqueness of an inbred, but also an evaluation of uniformity and stability, and in subsequent stages of inbred increase and hybrid seed production, an evaluation of genetic purity.

Determination of varietal status *de novo* according to the criteria of DUS is currently considered as different (and indeed a necessary prior determination) from any subsequent determination of essential derivation. Nonetheless, there exists the potential to greatly increase the efficiency of the EDV process; and ultimately therefore to improve the level of IP that is afforded to initial varieties. Quite simply, it should be possible to use the same set of SNPs that would be used as an initial step in the determination of Distinctness to also be the same set of markers that is used to help Essential Derivation. Individual PVP Offices could publish SNP profiles of all distinct varieties and thus individual breeders could utilize those data to also help obtain an initial determination of potential EDV status. There would then be no need for individual breeders to repeat the generation of these data and a current challenge in the initial determination of potential EDV status, not having access to the pertinent SNP profile of a proprietary inbred line bred by another breeder, would be removed.

2.5 Conclusions

Plant breeders who are employed by well-resourced agencies or companies increasingly have at their disposal technological capabilities to more effectively source useful genetics from a much broader base of diverse germplasm than was available, either to those who invented agriculture, or to generations of farmers or previous generations of plant breeders who have provided stewardship and gradually improved the performance of crop varieties. Whether plant breeders will actually use those genetic resources will depend upon the level of innovative research and product development that they can bring to bear in their breeding programs. The range of available IPP influences the range and type of research and product development that can be accomplished, at least by a private commercially funded organization. To have accessible a full range of choice in the level or type of IPP that is available allows breeders in turn a full range of choice to determine the level of innovativeness, risk taking, level and term of research investments that can be a sustainable business proposition for their enterprise. Exclusions in the level and type of IPP will lead to exclusions in the amount of innovation and risk taking that breeders will be able to exercise in the research and product development programs. Lowering the available ceiling of IPP will limit advances in productivity and in the absence of publicly funded support, reduce, or at least slow, access to a broader base of genetic diversity in breeding and in agriculture.

SNP data can be used to characterise genetic resources and they can also be used to establish and to validate the varietal status of new plant varieties. Using a SNP based system to characterise varieties it will be possible to make comparisons to all previously recognized varieties without the need to manage reference collections in

the field. There are opportunities for true harmonization of data and databases on a global basis by escaping the constraints that inevitably emanate due to the large GxE interaction effects associated with morphological characteristics. Most importantly, there are opportunities in the judicious use of sequence data to improve the level and efficiency by which IP is afforded through the PVP process.

The means to characterize and to determine eligibility of newly developed genotypes for varietal status and use in agriculture will one day finally catch up with the methods breeders use to help create those new genotypes. Such a development should result in a more efficient IP system, one that is simpler and more effective to police. The overall goals should be to help provide a business environment that will allow a greater breadth of genetic diversity to be surveyed with ever-increasing effectiveness and so to continually improve abilities to select new genotypes that are optimally able to perform in target agricultural environments. An appreciation of the urgency to radically and quickly, yet sustainably improve agricultural productivity can be comprehended by thinking back once again to the dawn of agriculture itself. As quoted by Clive James, the founder of the International Service of the Acquisition of Agri-biotech Applications, it has been estimated that “in the next 50 years, the global population will consume twice as much food as has ever been consumed since agriculture began 10,000 years ago.” (Arabic Knowledge @ Wharton 2012; Hoisington et al. 1999). Nearly one-third of this 50 year forecast time span has already been spent.

References

- Ammann K (2008) Integrated farming: why organic farmers should use transgenic crops. *N Biotechnol* 25:101–107
- Ammann K (2009) Why farming with high tech methods should integrate elements of organic agriculture. *N Biotechnol* 26:378–388
- Arabic Knowledge@Wharton (2012) Can biotechnology solve China’s food security problem? Wharton University of Pennsylvania. <http://knowledge.wharton.upenn.edu/arabic/article.cfm?articleid=2850>. Accessed 18 Oct 12
- Araus JL, Ferrio JP, Buxo R, Voltas J (2007) The historical perspective of dryland agriculture: lessons learned from 10,000 years of wheat cultivation. *J Exp Bot* 58:131–145
- Austin DF, Lee M, Veldboom LR (2001) Genetic mapping in maize with hybrid progeny across testers and generations: plant height and flowering. *Theor Appl Genet* 102:163–176
- Austin RB, Arnold MH (1989) Variability in wheat yields in England: analysis and future prospects. In: Anderson JR, Hazell PBR (eds) *Variability in grain yields implications for agricultural research and policy in developing countries*. Johns Hopkins University Press, Baltimore
- Bennett AJ, Bending GD, Chandler D et al (2011) Meeting the demand for crop production: the challenge of yield decline in crops grown in short rotations. *Biol Rev Camb Philos Soc* 87:52–71
- Bhatramakki D, Dolan M, Hanafey M et al (2002) Insertion–deletion polymorphisms in 3’ regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Mol Biol* 48:539–547
- Borlaug NE, Dowsell CR (2005) Feeding a world of ten billion people: a 21st century challenge. In: Tuberosa R, Phillips RL, Gale M (eds) *Proceedings of the International Congress: in the wake of the double helix: from the green revolution to the gene revolution*, 27–31 May 2003, Bologna, Italy. Avenue Media, Bologna, pp 3–23
- Bredemeijer GMM, Cooke RJ, Ganai MW et al (2002) Construction and testing of a microsatellite database containing more than 500 tomato varieties. *Theor Appl Genet* 105:1019–1026

- Brookes G, Barfoot P (2008) Global impact of biotech crops: socio-economic and environmental effects, 1996–2006. *AbBioForum* 11:21–38
- Calderini DF, Slafer GA (1998) Changes in yield and yield stability in wheat during the 20th century. *Field Crops Res* 57:335–347
- CAMBIA (undated) Can IP rights protect plants? Patent Lens. <http://www.patentlens.net/daisy/patentlens/1234.html>. Accessed 17 Oct 2012
- Castleberry RM, Crum CW, Krull CF (1984) Genetic improvement of U.S. maize cultivars under varying fertility and climatic conditions. *Crop Sci* 24:33–36
- Ching A, Caldwell KS, Jung M et al (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet* doi:10.1186/1471-2156-3-19
- Cohen JI (2000) Managing intellectual property: challenges and responses for agricultural research institutes. In: Persley GJ, Latin MM (eds) *Agricultural biotechnology and the poor: proceedings of an international conference*. CGIAR, Washington DC
- Crookston RK (2006) A top 10 list of developments and issues impacting crop management and ecology during the past 50 years. *Crop Sci* 46:2253–2262
- DEFRA (2009) The potential to increase productivity of wheat and oilseed rape in the UK. Report to the government chief scientific adviser. Dept. for the Environment, Food, and Regional Affairs, London
- Duvick DN (2005) The contribution of breeding to yield advances in maize (*Zea mays* L.). *Adv Agron* 86:83–145
- Enoki H, Miki K, Koinuma K (2006) Mapping of quantitative trait loci associated with early flowering of a northern flint maize (*Zea mays* L.) inbred line. *Maydica* 51:515–523
- ESA (2011) Position on Concept of EDV. ESA_11.0043. Eur Seed Assoc, Brussels
- FAOSTAT (2011) Statistics Office of FAO. <http://faostat.fao.org>. Accessed 18 Oct 2012
- Fernandez-Cornejo J (2004) The seed industry in U.S. agriculture: an exploration of data and information on crop seed markets, regulation, industry structure, and research and development. *Agric Inf Bull (U S Dep Agric)* No 786, Washington, DC
- Foley JA, Ramankutty N, Brauman KA et al (2011) Solutions for a cultivated planet. *Nature* 478:337–342
- Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci U S A* 99:9573–9578
- Glasmann JC, Kilian B, Upadhyaya HD, Varshney RK (2010) Accessing genetic diversity for crop improvement. *Curr Opin Plant Biol* 13:167–173
- Godfray HCJ (2011) Food and Biodiversity. *Science* 333:1231–1232
- Green RE, Cornell SJ, Scharlemann JPW, Balmford A (2005) Farming and the fate of wild nature. *Science* 307:550–555
- Hayes DJ, Lence SH, Goggi S (2009) Impact of intellectual property rights in the seed sector on crop yield growth and social welfare: a case study approach. *AgBioForum* 12:155–171
- Heckenberger M, Bohn M, Frisch M et al (2005a) Identification of essentially derived varieties with molecular markers: an approach based on statistical test theory and computer simulations. *Theor Appl Genet* 111:598–608
- Heckenberger M, Bohn M, Klein D, Melchinger AE (2005b) Identification of essentially derived Varieties obtained from biparental crosses of homozygous lines: II. Morphological distances and heterosis in comparison with simple sequence repeat and amplified fragment length polymorphism data in Maize. *Crop Sci* 45:1132–1140
- Heckenberger M, Bohn M, Melchinger AE (2005c) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines: I. Simple sequence repeat data from maize inbreds. *Crop Sci* 45:1120–1131
- Hof IL, Reid A (2008) Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue part I: discussion of morphological and molecular data (revised). 11th session of the working group on biochemical and molecular techniques and DNA profiling in particular, Madrid, Sept 16–18, 2008. BMT/11/0 Rev, UPOV, Geneva, Switzerland

- Hoisington D, Khairallah M, Reeves T et al (1999) Plant genetic resources: what can they contribute toward increased crop productivity? *Proc Natl Acad Sci U S A* 96:5937–5943
- ISF (2004a) Guidelines for the handling of a dispute on essential derivation in Lettuce. Int Seed Federa, Nyon, Switzerland
- ISF (2004b) Technical Protocol for Implementation of the ISF Guidelines for the Handling of a Dispute on EDV in Lettuce. Int Seed Federa, Nyon, Switzerland
- ISF (2005) Essential Derivation Information and Guidance to Breeders. Int Seed Federa, Nyon, Switzerland
- ISF (2006) Use of DNA markers for DUS testing, essential derivation and identification. Int Seed Federa, Nyon, Switzerland
- ISF (2007a) Guidelines for the handling of a dispute on essential derivation in cotton. Int Seed Federa, Nyon, Switzerland
- ISF (2007b) Guidelines for the handling of a dispute on essential derivation in oilseed rape. Int Seed Federa, Nyon, Switzerland
- ISF (2008) Guidelines for the handling of a dispute on essential derivation of maize lines. Int Seed Federa, Nyon, Switzerland
- ISF (2009) Guidelines for handling a dispute on essential derivation in ryegrass. Int Seed Federa, Nyon, Switzerland
- ISF (2012) ISF View on intellectual property. Int Seed Federa, Nyon, Switzerland. http://www.worldseed.org/cms/medias/file/PositionPapers/OnIntellectualProperty/View_on_Intellectual_Property_2012.pdf. Accessed 18 Oct 12
- JIC (2012) JIC statement on intellectual property, John Innes Centre, Norwich. <http://www.jic.ac.uk/corporate/about/policies/ip-policy.htm>. Accessed 18 Oct 2012
- Jones ES, Sullivan H, Bhatramakki D, Smith JS (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L.). *Theor Appl Genet* 115:361–371
- Jones H, Jarman RJ, Austin L et al (2003) The management of variety reference collections in distinctness, uniformity and stability testing of wheat. *Euphytica* 132:175–184
- Kahler AL, Kahler JL, Thompson SA et al (2010) North American study on essential derivation in Maize: II. selection and evaluation of a panel of simple sequence repeat loci. *Crop Sci* 50:486–503
- Kaufmann K, Pajoro A, Angenent GC (2010) Regulation of transcription in plants: mechanisms controlling developmental switches. *Nat Rev Genet* 11:830–842
- Krattiger AF (2004) Editor's introduction: PVP and agricultural productivity. *IP Strategy Today* 9:ii–vi
- Lai J, Li R, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42:1027–1030
- Lal R (2001) Managing world soils for food security and environmental quality. *Adv Agron* 74:155–192
- Laval G, SanCristobal M, Chevalet C (2002) Measuring genetic distances between breeds: use of some distances in various short term evolution models. *Genet Sel Evol* 34:481–507
- Law JR, Anderson SR, Jones ES et al (2011a) Approaches to improve the determination of eligibility for plant variety protection: I Evaluation of morphological characteristics. *Maydica* 56:1–18
- Law JR, Anderson SR, Jones ES et al (2011b) Approaches to improve the determination of eligibility for plant variety protection: II Identification and evaluation of a core set of morphological characteristics. *Maydica* 56:209–219
- Law JR, Anderson SR, Jones ES et al (2011c) Characterization of maize germplasm: comparison of morphological datasets compiled using different approaches to data recording. *Maydica* 56–1708. http://www.maydica.org/articles/56_069.pdf. Accessed 18 Oct 2012
- Le Buanec B (2004) Protection of plant-related innovations: evolution and current discussion. *IP Strategy Today* 9:1–18
- Li Y, Dong Y, Niu S, Cui D (2007) The genetic relationship among plant-height traits found using multiple-trait QTL mapping of a dent corn and popcorn cross. *Genome* 50:357–364

- Mackay I, Horwell A, Garner J et al (2011) Reanalyses of historical series of UK variety trials to quantify the contributions of genetic and environmental factors to trends and variability in yield over time. *Theor Appl Genet* 122:225–238
- Mackay TFC (2009) A-maize-ing Diversity. *Science* 325:688–689
- Malik S (2012) Food prices expected to rise after second wettest summer on record the guardian. <http://www.guardian.co.uk/environment/2012/oct/10/food-prices-rise-wettest-summer>. Accessed 10 Oct 2012
- Marlander B, Hoffmann C, Koch H-J et al (2003) Environmental situation and yield performance of the sugar beet crop in Germany: heading for sustainable development. *J Agron Crop Sci* 189:2012–2026
- Martienssen RA, Colot V (2001) DNA methylation and epigenetic inheritance in plants and filamentous fungi. *Science* 293:1070–1074
- Mauria S, Singh NN, Mukherjee AK, Bhat KV (2000) Isozyme characterization of Indian maize inbreds. *Euphytica* 112:253–259
- Mauria S, Singh NN, Bhat KV, Lakhanpaul S (2002) Assessment of genetic variation in Indian maize inbreds using RAPD markers. *J Genet Breed* 56:15–19
- Mickelson SM, Stuber CS, Senior L, Kaeppeler SM (2002) Quantitative trait loci controlling leaf and tassel traits in a B73 × Mo17 Population of Maize. *Crop Sci* 42:1902–1909
- MMEDV (1999) Molecular and other markers for establishing essential derivation in crop plants (EDV). EU-AgriNet. http://ec.europa.eu/research/agriculture/projects/qlrt_1999_01499_en.htm. Accessed 18 Oct 2012
- Nelson BK, Kahler AL, Kahler JL et al (2011) Evaluation of the numbers of single nucleotide polymorphisms required to measure genetic gain distance in maize (*Zea mays* L.). *Crop Sci* 51:1470–1480
- Nelson PT, Coles ND, Holland JB et al (2008) Molecular characterization of maize inbreds with expired U.S. Plant variety protection. *Crop Sci* 48:1673–1685
- NFU (2012) A mixed harvest, but wheat well down. National Farmers Union. <http://www.nfuonline.com/Your-sector/Crops/News/A-mixed-harvest,-but-wheat-well-down/>. Accessed 10 Oct 2012
- Ogilvie A, Farmer G (1997) Documenting the Medieval Climate. In: Hulme M, Barrow E (eds) *Climates of the British Isles: present, past and future*. Routledge, London
- Peng JH, Sun D, Nevo E (2011) Domestication evolution, genetics and genomics in wheat. *Mol Breed* 28:281–301
- Phalan B, Onial M, Balmford A, Green RE (2011) Reconciling food production and biodiversity conservation: land sharing and land sparing compared. *Science* 333:1289–1291
- Qin J, Chen W, Guan R et al (2006) Genetic contribution of foreign germplasm to elite chinese soybean (*Glycine max*) cultivars revealed by SSR markers. *Chin Sci Bull* 51:1078–1084
- Raven PH (2010) Does the use of transgenic plants diminish or promote biodiversity? *New Biotechnol* 27:528–533
- Rodrigues DH, de Alcantara Neto F, Schuster I (2008) Identification of essentially derived soybean cultivars using microsatellite markers. *Crop Breed Appl Biotechnol* 8:74–78
- Ronald P (2011) Plant Genetics, sustainable agriculture and global food supply. *Genet* 188:11–20
- Rudel TK, Schneider L, Uriarte M et al (2009) Agricultural intensification and changes in cultivated areas, 1970–2005. *Proc Natl Acad Sci U S A* 106:20675–20680
- Russell WA (1984) Agronomic performance of maize cultivars representing different eras of maize breeding. *Maydica* 29:375–390
- SGRP (2010) Booklet of CGIAR centre policy instrument, guidelines and statements on genetic resources, biotechnology and intellectual property rights. Version III. System-wide genetic resources program (SGRP) and the CGIAR genetic resources policy committee (GRPC). Bioiversity Int. Rome. http://www.sgrp.cgiar.org/sites/default/files/Policy_Booklet_Version3.pdf. Accessed 09 Oct 2012
- Smith BD (1989) Origins of agriculture in Eastern North America. *Science* 246:1566–1571

- Sourdille P, Baud S, Leroy P (1996) Detection of linkage between RFLP markers and genes affecting anthocyanin pigmentation in maize (*Zea mays* L.). *Euphytica* 91:21–30
- Tenaillon MI, Sawkins MC, Long AD et al (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci U S A* 98:9161–9166
- The Royal Society (2009) Reaping the benefits: Science and the sustainable intensification of global agriculture. ISBN: 978-0-8540-784-1. The Royal Society, London
- Tian F, Bradbury PJ, Brown PJ et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
- Van Inghelandt D, Melchinger AE, Lebreton C, Stich B (2010) Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theor Appl Genet* 120:1289–1299
- Vroh Bi I, McMullen MD, Sanchez-Villeda H et al (2006) Single nucleotide polymorphisms and insertion-deletions for genetic markers and anchoring the maize fingerprint contig physical map. *Crop Sci* 46:12–21
- Wang Q, Dooner HK (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the *bz* locus. *Proc Natl Acad Sci U S A* 103:17644–17649
- Warburton ML, Crossa J, Franco J et al (2006) Bringing wild relatives back into the family: recovering genetic diversity in CIMMYT improved wheat germplasm. *Euphytica* 149:289–301
- Williams SB, Weber KA (1989) Intellectual property protection and plants. In: Caldwell BE (ed) Intellectual property rights associated with plants. ASA Spec. Publ. No. 52. ASA, CSSA, and SSSA, Madison
- Yan J, Warburton M, Crouch J (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. *Crop Sci* 51:433–449
- Yu J, Zhang Z, Zhu C et al (2009) Simulation appraisal of the adequacy of number of background markers for relationship estimation in association mapping. *Plant Genome* 2:63–77

Chapter 3

The Use of Molecular Marker Data to Assist in the Determination of Essentially Derived Varieties

J. Stephen C. Smith, Elizabeth S. Jones and Barry K. Nelson

Contents

3.1	Introduction	50
3.2	Main Objectives for Introducing and Implementing the EDV Concept	52
3.2.1	Who Determines EDV Status?	53
3.2.2	Why Have Breeders Taken the Initiative to Help Determine What Constitutes an EDV?	53
3.2.3	How is EDV Status Determined?	54
3.3	Predominant Derivation	54
3.4	Measure of Conformity: A Clear Starting Point	55
3.5	The Use of Molecular Markers to Help Determine EDV Status	57
3.5.1	What Degree of Similarity is Required to Determine that a Variety is “Essentially Derived”?	58
3.5.2	Using Molecular Markers to Help Determine Essential Derivation in Maize: A Case Study	59
3.6	Ruling by the Court of Appeals, The Hague in <i>Danziger Flower Farm vs. Astee Flowers</i> on Technical Issues	60
3.7	Concluding Comments	61
	References	62

Abstract A primary reason to introduce the concept of an Essentially Derived Variety (EDV) was to retain the effectiveness of the *sui generis* Plant Variety Protection or Plant Breeders’ Rights system for the protection of new plant varieties in an era where it had become increasingly common and with greater facility to make relatively small

J. S. C. Smith (✉) · B. K. Nelson
DuPont Pioneer, 7300 NW 62nd Avenue, P.O. Box 1004, Johnston, Iowa 50131, USA
e-mail: stephen.smith@pioneer.com

E. S. Jones
Syngenta Biotechnology, Inc., 3054 East Cornwallis Road,
Research Triangle Park, Raleigh, NC 27709-2257
e-mail: liz.jones@syngenta.com

B. K. Nelson
e-mail: barry.nelson@pioneer.com

R. Tuberosa et al. (eds.), *Genomics of Plant Genetic Resources*,
DOI 10.1007/978-94-007-7572-5_3,

© Springer Science+Business Media Dordrecht 2014

genetic changes to existing varieties including through the use of transgenic modification, mutation breeding, and through the use in progeny selection using molecular markers. An initial concept was that EDV status would help provide a balanced approach to protection for the developer of the initial germplasm and a subsequent breeder who chose to make a change to that existing initial variety, one that may be genetically small, but with significant agronomic and economic consequences. The EDV concept can also help prevent plagiarism. Molecular markers are an important means to help determine EDV status. Crop specific guidelines have been developed. The technological and economic environment in which plant breeders operate will likely continue to change. These changes may well cause plant breeders and policy makers to seek to further adjust the *sui generis* intellectual property (IP) system so it can optimally contribute to and compliment other forms of intellectual property protection (IPP). The comprehensive goal should be to ensure that the various forms of IP that are available collectively maximize incentives to invest in the comprehensive range of breeding, research, and germplasm management activities that are required to develop improved varieties, both for today and for the long-term.

Keywords Plant Variety Protection (PVP) · Plant Breeders' Rights (PBR) · Essentially Derived Variety (EDV) · Molecular marker · Predominant derivation · Genetic conformity · Burden of proof · Intellectual Property Protection (IPP) · Morphology · Pedigree · Agronomic traits · Cosmetic breeding · Plagiarism · Genetic diversity · Germplasm · Transformation · Transgenic · Genetic engineering · Backcrossing · Reverse breeding · Mutation breeding · Double-Haploid

3.1 Introduction

Plant breeders create new combinations of germplasm in order to provide the genetic basis of new varieties with improved agronomic traits that are required by farmers, processors, or consumers. If currently available varieties are well adapted to the needs of farmers and consumers, breeders will seek to conserve most, if not all, of the favorable genetic combinations while continuing to introduce a sufficiency of appropriate germplasm to improve a few targeted traits (Troyer and Rocheford 2002). For example, the introduction of the semi-dwarf characteristic in wheat (*Triticum aestivum* L.) was a critically important component of the breeding strategy that drove the "Green Revolution," led by Dr. Norman Borlaug. And in the US, the introgression of just a couple of semi-dwarf genes (*Rht1* and *Rht2*, now known as *Rht-B1b* and *Rht-D1b*) which began in the 1950s and 1960s, greatly improved resistance to lodging, which facilitated the use of higher rates of nitrogen fertilizer and the subsequent genetic yield improvements in varieties that followed.

In addition, significant agronomic improvements can also be achieved through more radical genetic changes being made to existing widely used germplasm. Examples include: (1) the registration of the Pioneer brand maize hybrid "Dea" in France

(1980) which launched the “Iodent Revolution” (Barriere et al. 2006) and replaced older US-derived “flint” germplasm with US Corn Belt Dent germplasm and (2) the recent introduction into China of maize hybrids with improved agronomic performance at higher planting densities as a result of their very different germplasm constitution from other current widely used Chinese maize hybrids (Li et al. 2011). Similarly, in wheat (*Triticum aestivum* L.) relatively large genetic changes including the deployment of CIMMYT spring wheat germplasm into several regions around the world drove significant performance increases. Introductions of germplasm from CIMMYT spring wheat varieties that was hitherto not present in the US have played a major role in improving the performance of US wheat varieties. After three decades of breeding, approximately 12.5 % of the pedigree of some US wheat varieties traces to the more recently introduced CIMMYT spring wheat germplasm.

An essential strategy in plant breeding is to manage the amount and quality of allelic diversity from which new recombinants and segregants with improved genetic potential (genetic gain) can be selected. An important means to manage the amount of diversity is through the choice of breeding parents. Most diversity in a segregating breeding population will be created by crossing two or more parents (a synthetic or population) that are unrelated by pedigree. Least segregating diversity will be created by narrowing the breeding pedigree by crossing highly related parents including following multiple backcrosses of a recurrent parent, or as a result of adding a gene or stack of genes using transgenic technologies to an existing variety or inbred line. In addition, breeders can regulate the amount of diversity by controlling meiotic recombination and segregation: Diversity can be increased by undertaking generations of random mating or it can be decreased by the creation of haploids and the subsequent doubling of their genomes to recreate diploid progeny.

It has always been possible to develop a new plant variety that may be very similar in its morphological attributes and also in its genetic constitution and pedigree to a previously protected variety. For example, backcrossing is a breeding strategy that is pursued specifically for the purpose of recovering the majority of the genetic background and thus, the agronomic or horticultural features of the recurrent parent. The creation and selection of mutants or the selection of recoveries from inbred lines or varieties that were not thoroughly selected for homozygosity are additional means to develop new lines or varieties that are very similar to an existing variety or inbred line. Thus, while these new varieties may be distinct for one or a few morphological characteristics that constitute the set used to determine eligibility for PVP through tests for Distinctness, Uniformity, and Stability (DUS), they will nonetheless inevitably have a similar germplasm constitution, pedigree, and field performance attributes to the initial variety from which they were predominantly derived.

Advances in biotechnology include many new methods to create mutations and screen for their efficacy, to more rapidly develop homozygous inbred lines, to screen for homozygosity, to aid in selection and reduce the time required in backcrossing, and to introduce transgenic or native traits (Smith et al. 2008). These breeding strategies have, among their other uses, also increased the means and the facility by which closely similar, yet morphologically distinct varieties can be developed and protected

by PVP. The European Seed Association (ESA 2011) notes that “In the light of modern breeding techniques, it has become much more likely that a variety bred from an existing variety (initial variety) in its essential characteristics still conforms to the initial variety.” If changes to existing varieties involve a relatively small proportion of their germplasm, yet they contribute to improved agronomic features (e.g., by adding an important disease or insect resistance trait by conventional breeding or by transformation), it is potentially beneficial to farmers and consumers to allow commercialization of the new improved varieties. Alternatively, other small changes to the existing genotype may result in distinct phenotypic differences, but which are simply cosmetic (GHK 2011) since they do not improve agronomic features. In this latter event, the derivative can be considered a “me too” variety and so represents the outcome of plagiarism rather than productive plant breeding, or as described by van Eeuwijk and Law (2004), “fraudulent practices in which ‘new’ varieties are produced from current, protected ones without a genuine breeding effort.” The Essentially Derived Variety (EDV) concept provides a safeguard to the owner of an initial variety to protect the IP he, or she, has created in the event another breeder makes a relatively small genetic change to that variety. If that genetic change contributes a useful additional agronomic feature, the owner of the initial variety may find it advantageous to jointly commercialize the new variety with the second breeder under mutually agreed terms. Alternatively, if the initial breeder considers the change made by the second breeder to potentially undermine his rights then he can refuse to allow the EDV to be commercialized and so prevent plagiarism (ESA 2011).

3.2 Main Objectives for Introducing and Implementing the EDV Concept

Such rationale led to the introduction of the concept of an EDV (UPOV 1991). For example, a recent review of the European Community Plant Variety Right (CPVR) (GHK 2011) states: “An important implication of the breeders’ exception is that a variety that is only marginally different from a protected variety could qualify for protection as a new variety. Production of such ‘mimic’ varieties could deprive the original breeder of royalties from the protected variety. The ‘essentially derived variety’ provision in UPOV and repeated in the CPVR Basic Regulation is an attempt to reduce the problems with imitation that can result from the breeders’ exemption.”

In its broadest context, the objective of the EDV concept is to provide an equitable balance in returns from commercialization between the breeder of an initial variety and a subsequent developer who makes a relatively minor genetic change to the initial variety, but one which contributes to significant agronomic improvement of the initial variety. In these circumstances, there are potential benefits to society by allowing both breeders to share in the IP, and thus the financial benefits that accrue from commercialization of the subsequently developed variety.

3.2.1 Who Determines EDV Status?

UPOV purposely provides no specific guidance as to how EDV status should be determined. Such determinations have been left to experts in the relevant fields. And, if determinations cannot be agreed among breeders, a dispute resolution mechanism is available (ISF [undated a](#)). Ultimately, there is redress to the courts, and rulings will be made that may contribute to future precedence (District Court [2008](#); Court of Appeal [2009](#); GHK [2011](#)). GHK ([2011](#)) further notes that “the EDV provision is appropriate but the definition is unclear and there are few established protocols for making EDV determinations. There is scope for improvement in this area.”

3.2.2 Why Have Breeders Taken the Initiative to Help Determine What Constitutes an EDV?

ISF ([2005](#)) notes that “this principle (EDV) mainly involves questions of scope of protection and enforcement of the rights of the breeder.” It is left to the initiative of the breeder to enforce these rights. ISF stresses that “the determination of essential derivation is not part of the procedure of the granting of the Breeder’s Right.” “With regard to establishing whether a variety is an essentially derived variety, a common view expressed by members of the UPOV is that the existence of a relationship of essential derivation between protected varieties is a matter for the holders of plant breeders’ rights in the varieties concerned” (UPOV [2009](#)). The European Union Community Plant Variety Office also acknowledges that “In an application procedure, there is no role for authorities charged with granting plant variety rights to determine whether a variety is an EDV” (Kiewiet [2006](#)).

Breeders understand that a primary incentive to introduce the concept of EDV is to facilitate the continued improvement of existing varieties *via* the breeder exception of the PVP Act. With regard to EDV, the incentive is specifically to provide an equitable IP balance between the breeder of the initial variety and a subsequent breeder who makes an improvement to that variety while retaining the essential genetic basis, morphological appearance and agronomic performance attributes of that initial variety. In such circumstances, the breeder of the EDV most likely chose the initial variety as the genetic foundation upon which to make an agronomic improvement because that initial variety had performance attributes that breeder largely wanted to retain. Consequently, the new improved variety is dependent upon both breeders for its creation and so both deserve the opportunity for equitable remuneration.

Much discussion of how to interpret cases of essential derivation have occurred among breeders within the International Seed Federation (ISF) and assuredly also within national seed associations. The main gist of these discussions runs as follows: Breeders value access to an IP process that is simple, predictive, and that will not require diversion of resources or delays that protracted legal disputes can consume. Therefore, breeders considered it would be advantageous if they, together with other appropriate technical experts, would create guidelines on how to determine EDV status. It was feared that the absence of guidelines would serve only to increase

prospects for litigation, increase the resources consumed by such litigation, and increase prospects that legal precedents might be set that could be contrary to the best interests of encouraging productive plant breeding. It was felt that most meaningful guidelines would be developed by those who best understand the goals and practices of plant breeding and the diversity of germplasm that is available and routinely used in the specific crop species with which they are familiar.

It is perhaps not surprising to learn that EDV disputes involving two crop species, which did not have guidelines previously developed, have now been long argued. An EDV case in wheat (*Triticum aestivum* L.), after a succession of court hearings and multiple presentations of laboratory data has now reached the High Court after 10 years (UPOV BMT 2011). And after eight years in litigation, The Court of Appeal in The Hague rendered its decision regarding Danziger Flower Farm vs. Astee Flowers where the plaintiff alleged that the variety Blancanieves (*Gypsophila* spp.) was an EDV of the plaintiff's variety Dangypmini. The court ruled that Blancanieves was not an EDV of Dangypmini and made several rulings that touch upon both the legal and technical knowledge required to determine EDV status (which we will return to later in this chapter).

3.2.3 *How is EDV Status Determined?*

Article 14(5)(b) of the 1991 Act of the UPOV Convention states that “a variety shall be deemed to be essentially derived from another variety (“the initial variety”) when it is predominantly derived from the initial variety, or from a variety that is itself predominantly derived from the initial variety, while retaining the expression of the essential characteristics that result from the genotype or combination of genotypes of the initial variety, it is clearly distinguishable from the initial variety and except for the differences which result from the act of derivation, it conforms to the initial variety in the expressions of the essential characteristics that result from the genotype or combination of genotypes of the initial variety.”

There is a two-stage process: First, a new variety is submitted to testing for the DUS criteria required to obtain a PVP certificate. Second, determination of EDV status can occur. Assessment of essential derivation should consider the following requirements:

- Predominant derivation from the initial variety
- Conformity to the initial variety in the expression of the essential characteristics that result from the genotype or the combination of genotypes of the initial variety.

3.3 **Predominant Derivation**

The UPOV Convention (UPOV 1991) does not provide clarification of the term “predominantly derived” (UPOV 2009).

Predominant derivation from the initial variety implies that the initial variety or products essentially derived from the initial variety have been used in the breeding process. In order to prove that use, various criteria or a combination thereof may be used:

- Combining ability
- Phenotypic characteristics
- Molecular characteristics
- Breeding records

To address the issue of genetic distance measured between the initial variety and the putative EDV.

ISF (2005) notes that: “Any conventional breeding method could, in theory, provide an essentially derived variety.” UPOV provides additional clues that can be used to help address the issue of predominant derivation through the documentation of a non-exhaustive list of methods, which might lead to essential derivation. For example, Article 14 (c) of UPOV (1991) states: “Essentially derived varieties may be obtained for example, by the selection of a natural or induced mutant, or of a somaclonal variant, the selection of a variant individual from plants of the initial variety, backcrossing, or transformation by genetic engineering.” The Convention clarifies that these are examples and do not exclude the possibility of an EDV being obtained in other ways. For example, another way to obtain an EDV from an initial variety “could be the use of a hybrid variety to obtain a variety which is essentially derived from one of the parent lines of the hybrid” (UPOV 2009). In addition, predominant derivation could also occur from “the use of molecular marker data, of an initial variety, for the purpose of selection of genotypes very close to the genotype of its parental line(s) or of the initial hybrid itself” (ISF 2012).

3.4 Measure of Conformity: A Clear Starting Point

The owner of the initial variety will usually be disadvantaged in respect of being able to obtain the information required to determine EDV status. Therefore, ISF insists on the necessity of clearly defining a starting point in determining dependence or conformity (ISF 2005).

Of the evidential sources that can contribute to a determination of essential derivation, combining ability and phenotypic characteristics both require replicated field testing that is very consuming of time and resources. Furthermore, if the issue of EDV status involves proprietary parental inbred lines then those are usually maintained as proprietary property and consequently not freely available to the owner of the initial variety. Pedigree data of the suspected EDV may also not be readily available, or could be in error. Consequently, molecular data are sometimes the only initial source of evidential data that can be available to allow comparisons to be made by the owner of the initial variety between that initial variety and the putative EDV.

Even where other evidential data can be collected, marker data are still the least expensive to obtain and can be made available within a short timeframe. Once again, if the putative EDV is a parental line of a hybrid then comparisons can be complicated by the inability of the owner of the initial variety to access seed of parental lines that are maintained as trade secrets. In some circumstances, however, marker profiles of inbred parents of single cross hybrids can be deduced provided maternal (pericarp) tissue is available on the hybrid seed (Wang et al. 2002).

Therefore, to have a clear starting point to the process of determining EDV status the International Association of Plant Breeders (ASSINSEL) (now constituted within the International Seed Federation) argues that there must be a provision for the reversal of the burden of proof when the breeder of a pre-existing protected variety has been able to show that a new variety is potentially essentially derived from this pre-existing initial variety. This rationale is founded upon the fact that ultimate proof of EDV status using data from all of the multiple sources of evidence such as combining ability, phenotypic characteristics, molecular characteristics and breeding records) would be impossible to be collected and assessed by the owner of the initial variety because most of these sources of information would be held by the developer of the putative EDV. Therefore, in the case of a variety that is very similar to the initial variety, it seemed fair and reasonable that it should be the developer of the putative EDV who then needs to prove that the disputed variety has not been essentially derived from the owner of the initial variety. ISF (2012) states: "It can be very challenging for the owner of the initial variety to prove predominant derivation. Consequently, ISF strongly believes that it is necessary for breeders to have the capability to reverse the burden of proof, so that it is then placed upon the breeder of the putative EDV, when a high degree of phenotypic and/or genetic conformity between the initial variety and the putative EDV has been established. If the owner of the initial variety has been able to show convincingly that the conformity requirement is fulfilled, the owner of the putative EDV will then have to prove that there is no predominant derivation; i.e., that he has not predominantly used the initial variety or a variety essentially derived from the initial variety". The European Seed Association also "supports the reversal of the burden of proof in favour of the holder of the plant breeders' right of the initial variety once a certain degree of genotypic similarity between the initial variety and a suspected essentially derived variety is reached." (ESA 2011).

In *Van Zanten Plants B.V. v. Hofland B.V.*, the District Court of The Hague (2008) stated that although "DNA analysis is not a requirement to obtain a plant variety right; it is not relevant to assess the criterion of distinctness. However, the results of such a DNA analysis constitute an important indication that there has been an act of (essential) derivation." (District Court 2008).

UPOV intends to leave the matter of reversal of burden of proof to Member States (Court of Appeal 2009). The Court of Appeals (2009) in *Danziger vs. Astee Flower farm* noted that under Dutch procedural law "(virtually) the same result can be achieved as with the reversal of the burden of proof, by taking the ground, based on evidence furnished by the breeder of the initial variety, that evidence of an EDV has been furnished for the time being and allowing the other party to prove the contrary."

3.5 The Use of Molecular Markers to Help Determine EDV Status

Molecular markers were first used in the plant breeding industry in the early 1980s. Comparisons of isoenzymes and seed storage protein profiles could be used to uniquely identify, for example, 85–90% of US maize inbred lines. They could be used to validate many pedigrees and they have been widely used to monitor genetic purity in maize during the last 30 years. However, abilities to measure genetic distances that are reflective of pedigree, especially when those pedigrees are closely related, requires the use of marker technologies that allow the genome to be assayed in much greater detail and completeness than do isozymes or seed storage protein loci; indeed in some species with a very narrow germplasm base (e.g., Cucumber or *Cucumis sativus* L.) then DNA sequence or Single Nucleotide Polymorphism (SNP) data are necessary (Staub et al. 2005). Determination of EDV can also be particularly challenging in crops where there are generally high levels of shared genetic background. Such examples include *Calluna vulgaris* L. (Hull.) (Borchert et al. 2008) and durum wheat *Triticum durum* (Desf.) (Maccaferri et al. 2007).

By the early 1990s, there was already a voluminous scientific literature showing that molecular markers such as Restriction Fragment Length Polymorphisms (RFLPs) could provide the basis for measuring genetic distances between varieties that were reflective of pedigree relatedness (Smith and Smith 1989; 1991; Melchinger et al. 1991; Smith et al. 1990, 1991; Messmer et al. 1993; Hahn et al. 1995; Plaschke et al. 1995). In contrast, researchers had found that distance coefficients measured using morphological data could have only limited power to distinguish between initial varieties and putative EDVs (Gilliland et al. 2000; Roldan-Ruiz et al. 2001) and they were not always reflective of pedigree or genetic relationships (Bar-Hen et al. 1995; Burstin and Charcosset 1997; Dillmann et al. 1997; Dillmann and Guerin 1998; Smith and Smith 1989; Lefebvre et al. 2001; Gunjaca et al. 2008; CPV5766 2008). In addition, it was realized that use of morphological data would be time consuming and expensive to obtain, and be vulnerable to genotype x environment effects. Consequently, “it was obvious that molecular markers would form the preferred way of establishing genetic conformity between varieties; as molecular markers reflect the genotype directly and do not require the time consuming field activity” (van Eeuwijk and Law 2004). Likewise, Heckenberger et al. (2005a) noted that “Because molecular markers, such as simple sequence repeats (SSRs) or amplified fragment length polymorphisms, allow tracing chromosomal segments from parents to their progeny, genetic similarities based on molecular markers were regarded as suitable tools to distinguish EDVs from independently derived varieties.” Furthermore, it would not be possible using morphological comparisons of hybrids to deduce and thus to compare the constitutions of parental lines whereas such capabilities, at least in some circumstances, can be afforded by the use of molecular markers (Wang et al. 2004). Heckenberger et al. (b) concluded that “morphological traits and heterosis are less suited for identification of EDV’s in maize than molecular markers.” Similarly, Rodrigues et al. (2008) concluded that molecular markers are more discriminative and thus more suitable than morphology to determine EDV status in soybean [*Glycine max* (L.) Merrill].

The European Union funded a research project entitled “Molecular and other Markers for Establishing Essential Derivation (EDV) in Crop Plants—MMEDV” to “assess the degree to which different marker technologies will be efficient, precise and suited to measuring the diversity between well-defined genotypes” (Leigh et al. 2005). As an example of the research sponsored by the EU, a report (MMEDV 1999) states: “We have unequivocally established that EDV can be measured using molecular markers. Moreover we have developed methods for its assessment in the three crops (*Rosa* spp., barley, and maize) under examination. We have developed a number of statistical tools for use in the measurement of EDV and example sets of background data for comparison. In a more generic sense this project has established a framework for the assessment of EDVs in any crop, giving clear principles of approach to the use of various analytical techniques. It is clear that morphological measures are generally inappropriate.”

These researchers have identified numerous technical issues that must be addressed regarding the use of molecular markers to help determine EDV status with regard to a specific crop species. These include: sampling of varieties, the effect of heterogeneity, the marker system(s) to be used, the number of markers, degree of genomic coverage, the ability of the markers to distinguish between varieties known to be different, the ability of markers to show associations that reflect known pedigrees, selection of varieties to constitute a reference set of known pedigrees, statistical methods to measure distances (ISF undated a). It is not possible in this chapter to provide a detailed review of these subjects. Therefore, readers are directed to Bernardo and Kahler (2001), Heckenberger et al. (2003, 2005a, b, c), Leigh et al. (2005), van Eeuwijk and Baril (2004), and van Eeuwijk and Law (2004). As noted by the EU study into EDVs, “the framework for the assessment of EDVs in any crop, giving clear principles of approach” has been achieved (MMEDV 1999).

3.5.1 What Degree of Similarity is Required to Determine that a Variety is “Essentially Derived”?

ISF (2006) states that: “DNA markers may also be used to define genetic similarity trigger points for starting a dispute resolution process in cases of alleged essential derivation.” The species-specific nature of many of the issues underlying the determination of marker based trigger points include breeding practice and the breadth of well adapted genetic diversity. The complexity of these and other important factors collectively argue for a crop -by-crop approach in order to determine technical guidelines. Such guidelines should include protocols describing molecular marker methods, how data are analyzed and rendered into pair-wise distance data among varieties, and ultimately, how similarity thresholds are determined. The latter can then contribute to the assessment of potential EDV status and provide evidence to address the legal question of whether there has been predominant derivation. The ISF provided EDV guidelines for Cotton (*Gossypium hirsutum*, ISF 2007a), lettuce (*Lactuca sativa*, ISF 2004a; 2004b), maize, (*Zea mays*, ISF 2008), oilseed rape (*Brassica rapa*, ISF 2007b) and ryegrass (*Lolium perenne* L., ISF 2009). In addition to the

crop specific guidelines which include details concerning the criteria previously discussed ISF also provides procedures for arbitration of disputes concerning essential derivation, technical rules for establishing a threshold of essential derivation, and a list of international arbitrators (ISF [undated b](#)).

3.5.2 Using Molecular Markers to Help Determine Essential Derivation in Maize: A Case Study

The American Seed Trade Association, through its Cultivar Variety Identification Sub-Committee, began to examine the technical issue of maize EDVs in the early 1990s. The initial methodology was to examine genetic distances between pairs of progeny and their parents developed from a range of breeding schemes. Two initial studies used RFLP data to examine genetic distances between progeny already developed by various breeding schemes and different companies. These studies were reported upon by Bernardo and Kahler ([2001](#)). Subsequently, specific breeding populations were developed that included derivation from the F2 generation and derivation from backcross generations. Progeny were again profiled using RFLPs and test cross data were examined. Additional studies were also concurrently underway in France and Germany. The French Maize Breeders Association (SEPROMA) used RFLPs to profile inbred lines that had already been awarded PVP certificates and so were distinct according to UPOV guidelines. However, some of these pairs of inbred lines were known from morphological, pedigree data (known only to companies who had bred those lines), and agronomic data were already suspected to be eligible as EDVs. The German study used simulated data that were validated by the production of parental-progeny triplets to help determine technical details of how to measure genetic distances.

Participants from all three studies met under the auspices of the ISF and agreed the following thresholds using Rogers distance measures of RFLP profile data (UPOV BMT [2007](#), UPOV BMT Add [2007](#)):

- Red zone: above 90 % of similarity
- Orange zone: between 90 and 85 %
- Green zone: below 85 %

Subsequently, during the mid-late 1990s, the relatively cumbersome RFLP technology was replaced by SSR technology. New analyses using microsatellites were completed by both ASTA and by SEPROMA. New thresholds were agreed upon to take into account the greater variability expressed by SSRs compared to RFLPs:

- Red zone: above 90 % of similarity
- Orange zone: between 90 and 82 %
- Green zone: below 82 %

A set of SSR loci has been selected by the French Association of Maize Breeders and which has been validated by ISF to help in the determination of EDV in maize

(Heckenberger et al. 2003). Likewise, the ASTA has published a set of SSR loci (ASTA undated) that can be used to help determine EDV status (Kahler et al. 2010).

In addition, a code of conduct using SSRs and these thresholds was adopted by members of the French maize seed industry as follows: “Above the threshold of 90 % the variety should be considered as an EDV without further discussion; between 82 and 90 % there is possible essential derivation and the parties have to negotiate; below 82 % there is no essential derivation.” (ISF 2005). ISF (undated c) has published explanatory notes for the arbitration of disputes concerning essential derivation including “an EDV threshold that forms the trigger point for the reversal of the burden of proof.”

Most recently, in 2010, a new EDV study was jointly initiated by the ASTA and the Union Française des Semenciers (UFS) (French Seed Producers), previously known as SEPROMA, taking advantage of the availability of a chip developed by Illumina, which could be used to assay tens of thousands of Single Nucleotide Polymorphisms (SNPs). Objectives of this study are to 1) translate agreed SSR-based EDV thresholds to SNP-based thresholds, 2) to determine the number of SNPs that are required to provide determinative evidence of genetic conformity, and 3) to identify a set of publicly available SNPs that can be used to help determine EDV status in maize. It is hoped that this set of SNPs can also be used as a starting point for maize breeders in other regions of the world to evaluate SNPs as the basis for helping to determine EDV status that will be relevant for the germplasm that they utilize.

3.6 Ruling by the Court of Appeals, The Hague in Danziger Flower Farm vs. Astee Flowers on Technical Issues

The Dutch Court of Appeals (Court of Appeal 2009) concluded that “the determination of genetic conformity between plant varieties by means of Amplified Fragment Length Polymorphism (AFLP) markers is open to objections.” The court highlighted the importance of using multi-allelic markers and reliably sampling the entire genome. Concerns expressed by the court were that 1) “dominant markers such as AFLPs overestimate the real degree of identity between genotypes” and 2) “it is unknown to what extent the markers . . . represent the *Gypsophila* genome.” The court also remarked upon the importance of providing an estimate of the reliability of the distance or similarity measures through the calculation of a standard error. With regard to these comments by the Court of Appeals it is important to reiterate that there were no breeder-derived and approved guidelines for the use of molecular marker data to help determine EDV status in *Gypsophila*. Experience with developing guidelines in other species indicates that these and other issues had already been taken into consideration. Concerns expressed by the Court about the use and interpretation of molecular marker data could have been addressed by a study of comparing varieties of *Gypsophila* of known pedigree, including some that were indeed regarded as EDVs.

3.7 Concluding Comments

The EDV concept was introduced in order to encourage both the creation of new improved (initial) varieties that are developed as a result of relatively significant changes in the genome compared to their parents and the making of relatively small genetic changes provided they also contributed significant agronomic improvements. It has always been possible to make relatively small genetic changes to existing plant varieties. Some of these relatively small genetic changes can materially contribute to improved agronomic performance and should be encouraged. However, under the 1978 *sui generis* system of UPOV, such relatively small genetic changes to a variety would result in a distinct new variety, which would then cause all the ownership rights of the initial variety to be captured by the second breeder. The economic outcome of such a situation would be to potentially eclipse the ability of the owner of the initial variety to obtain an economic return on his initial investments. In such a technological and economic environment who then would choose to invest in a subsequent cycle of developing a new variety which might only have too few years of commercial life to recoup the investments which lead to its development?

Consequently, just as plant breeders seek to develop new varieties which are better adapted to changing agro-ecological conditions, so then it was imperative that the IP system also be adapted to fit the economic and IP environment which had been altered by biotechnology. *Sui generis* IP systems, such as PVP are specially written for their subject matter, their purpose, and the economic and technical environment. With the advent of biotechnology, that technical environment had changed. It was considered imperative to then adjust the *sui generis* system. For without such a change, all breeders of initial varieties would be relatively disincentivised to continue the development of new improved initial varieties in an environment where others could be incentivised to conduct cosmetic breeding or to plagiarise existing varieties. As a consequence, the EDV concept was introduced with the objective to create an improved balance between breeders of initial varieties and later generation improvers of that germplasm. To estimate its effectiveness at achieving these objectives, according to the results of a stakeholder consultation survey carried out in the EU: “Most respondents emphasized that the EDV provision discourages ‘plagiarism’ of varieties and facilitates research and investment in breeding activity.” (GHK 2011). This report states that “Disagreements over EDV determination can be difficult to resolve where there are no established procedures or thresholds, and industry would benefit from these, particularly in court procedures.” GHK (2011) further states: “Enforcement is essential to a rights holders’ ability to effectively protect and exploit their invention. Effective enforcement is essential to incentivising innovation in agriculture for the EU.”

Plant breeders increasingly understand the genetic basis of many important agronomic traits, including those that are under quantitative genetic control (Tuberosa and Salvi 2009). An increasingly knowledge-based genetic approach to plant breeding can facilitate the identification and incorporation of useful new genetic diversity (Tanksley and McCouch 1997). Breeders will be able to more effectively target specific agronomic traits for improvement (Barriere et al. 2006). The EDV concept

might be anticipated to play an increasingly important role in cultivar improvement as additional traits, or at least important genetic components controlling those traits, are discovered and rapidly introgressed into existing varieties.

It is probably on balance advantageous to have sound technical guidelines which can provide clarity to helping to determine EDV and predominant derivation (Kiewiet 2006; GHK 2011). Alternatives include the possibility of competing and contrary decisions from courts which then fail to set clear precedence and serve to further prolong litigation. Defining marker thresholds for EDV on an individual crop basis greatly assists in providing clarity to owners of the initial variety, competitor breeders seeking to utilize the initial variety in their breeding programs, as well as the legal system. However, facilitated misuse of the EDV system is also a possibility. For example, a breeder might deliberately maintain the majority of the genome of an initial variety that contributed to its agronomic performance while also selecting against marker similarity to the initial variety at other genomic regions with little agronomic effect so that the resultant genotype would be below the EDV threshold. If such activities were to occur then they could undermine the basic tenets upon which the concept of an EDV was initially founded. Consequently, users of the EDV system should seek to continually address the guidelines to ensure that they are not being used to the detriment of crop diversity and agronomic value.

The success of any IP system in providing an environment that encourages the development of varieties with improved agronomic performance and increased productivity will always be dependent upon how the IP instrument is adapted to the economic and technological environment. Breeders and policy makers should continually review the capabilities of available IP instruments. The technological and economic environment in which plant breeders operate will likely continue to change. These changes may well cause plant breeders and policy makers to seek to further adjust the *sui generis* IP system so it can optimally contribute to and compliment other forms of IPP. The comprehensive goal should be to ensure that the various forms of IP that are available collectively maximise incentives to invest in the comprehensive range of breeding, research, and germplasm management activities that are required to develop improved varieties, both for today and for the long-term.

References

- ASTA (undated) Simple sequence repeat markers. Am seed trade assoc. Alexandria www.amseed.org/news_srr.asp. Accessed 22 Oct 2012
- Bar-Hen A, Charcosset A, Bougoin M, Guiard J (1995) Relationship between genetic markers and morphological traits in a maize inbred lines collection. *Euphytica* 84:145–154
- Barriere Y, Alber D, Dolstra O et al (2006) Past and prospects of forage maize breeding in Europe: II. History, germplasm evolution, and corroborative agronomic changes. *Maydica* 51:435–449
- Bernardo R, Kahler AL (2001) North American study on essential derivation in maize: inbreds developed without and with selection from F2 populations. *Theor Appl Genet* 102:986–992
- Borchert T, Krueger J, Hohe A (2008) Implementation of a model for identifying essentially derived varieties in vegetatively propagated *Calluna vulgaris* varieties. *BMC Genet* doi:10.1186/1471-2156-9-56

- Burstin J, Charcosset A (1997) Relationship between phenotypic and marker distances: theoretical and experimental investigations. *Heredity* 79:477–483
- Court of Appeals (2009) Judgement of the seventh civil court decision of 29 December 2009. In the case of DANZIGER “DAN” FLOWER FARM vs. ASTEE FLOWERS B.V. Court of Appeals in the Hague, The Hague, Netherlands, 15pp
- CPV5766 (2008) Final report: Management of winter oilseed rape reference collections. National institute of agricultural botany (on behalf of the community plant variety office), Cambridge
- Dillmann C, Bar-Hen A, Guerin D et al (1997) Comparison of RFLP and morphological distances between maize (*Zea mays* L.) inbred lines. Consequences for germplasm protection purposes. *Theor Appl Genet* 95:92–102
- Dillmann C, Guerin D (1998) Comparison between maize inbred lines: genetic distances in the expert’s eye. *Agronomie* 18:659–667
- District Court (2008) Van Zanten Plant BV v. Hofland BV summary. The Hague, The Netherlands
- ESA (2011) Position on concept of EDV. ESA_11.0043. *Eur Seed Assoc.* http://www.euroseeds.org/publications/position-papers/intellectual-property/esa_11.0043. Accessed 22 Oct 2012
- GHK (2011) Evaluation of the community plant variety right *Acquis*—Final report. GHK, London
- Gilliland TJ, Coll R, Calsyn E et al (2000) Estimating genetic conformity between ryegrass (*Lolium*) varieties. I. Morphology and biochemical characterization. *Mol Breed* 6:569–580
- Gunjaca J, Buhinicek I, Jukic M et al (2008) Discriminating maize inbred lines using molecular and DUS data. *Euphytica* 161:165–172
- Hahn V, Blankenhorn K, Schwall M, Melchinger AE (1995) Relationships among early European maize inbreds: III. Genetic diversity revealed with RAPD markers and comparisons with RFLP and pedigree data. *Maydica* 40:299–310
- Heckenberger M, van der Voort JR, Melchinger AE et al (2003) Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties: II. Genetic and technical sources of variation in AFLP data and comparison with SSR data. *Mol Breed* 12:97–106
- Heckenberger M, Bohn M, Frisch M et al (2005a) Identification of essentially derived varieties with molecular markers: An approach based on statistical test theory and computer simulations. *Theor Appl Genet* 111:598–608
- Heckenberger M, Bohn M, Klein D, Melchinger AE (2005b) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines: II. Morphological distances and heterosis in comparison with simple sequence repeat and amplified fragment length polymorphisms. *Crop Sci* 45:1132–1140
- Heckenberger M, Bohn M, Melchinger AE (2005c) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines: I. Simple sequence repeat data from maize inbreds. *Crop Sci* 45:1120–1131
- ISF (2004a) Guidelines for the handling of a dispute on essential derivation in lettuce. Int Seed Federa, Nyon
- ISF (2004b) Technical protocol for implementation of the ISF guidelines for the handling of a dispute on EDV in lettuce. Int Seed Federa, Nyon
- ISF (2005) Essential derivation information and guidance to breeders. Int Seed Federa, Nyon
- ISF (2006) Use of DNA markers for DUS testing, essential derivation and identification. Int Seed Federa, Nyon
- ISF (2007a) Guidelines for the handling of a dispute on essential derivation in Cotton. Int Seed Federa, Nyon
- ISF (2007b) Guidelines for the handling of a dispute on essential derivation in Oilseed rape. Int Seed Federa, Nyon
- ISF (2008) Guidelines for the handling of a dispute on essential derivation of Maize lines. Int Seed Federa, Nyon
- ISF (2009) Guidelines for handling a dispute on essential derivation in Ryegrass. Int Seed Federa, Nyon

- ISF (2012) ISF View on intellectual property. Int Seed Federa, Nyon. http://www.worldseed.org/cms/medias/file/PositionPapers/OnIntellectualProperty/View_on_Intellectual_Property_2012.pdf. Accessed 22 Oct 2012
- ISF (undated a) Issues to be addressed by technical experts to define molecular marker sets for establishing thresholds for ISF EDV arbitration. Int Seed Federa, Nyon. www.worldseed.org/cms/medias/file/Rules/EssentialDerivation/Threshold_ISF_EDV_Arbitration.pdf. Accessed 22 Oct 2012
- ISF (undated b) Essential derivation. Int Seed Federa, Nyon. <http://www.worldseed.org/isf/edv.html>. Accessed 22 Oct 2012
- ISF (undated c) Explanatory notes: Regulation for the arbitration of disputes concerning essential derivation (RED). Int Seed Federa, Nyon. http://www.worldseed.org/cms/medias/file/Rules/EssentialDerivation/Explanatory_Notes.pdf. Accessed 22 Oct 2012
- Kahler AL, Kahler JL, Thompson SA et al (2010) North American study on essential derivation in Maize: II. Selection and evaluation of a panel of simple sequence repeat loci. *Crop Sci* 50:486–503
- Kiewiet B (2006) Essentially derived varieties. Community plant variety office, European Union, Angers
- Lefebvre V, Goffinet B, Chauvet JC et al (2001) Evaluation of genetic distances between pepper inbred lines for cultivar protection purposes: comparison of AFLP, RAPD and phenotypic data. *Theor Appl Genet* 102:741–750
- Leigh FJ, Law JR, Lea VJ et al (2005) A comparison of molecular markers for diversity and EDV assessments. In: Tuberosa R, Phillips RL, Gale M (eds) *Proceedings of the international congress in the wake of the double helix: From the green revolution to the gene revolution (27–31 May 2003)*. University of Bologna, Italy
- Li Y, Ma X, Wang T et al (2011) Increasing maize productivity in China by planting hybrids with germplasm that responds favorably to higher planting densities. *Crop Sci* 51:2391–2400
- Maccafferri M, Sanguineti MC, Xie C et al (2007) Relationships among durum wheat accessions: II. A comparison of molecular and pedigree information. *Genome* 50:385–399. doi:10.1139/G07-017
- Melchinger AE, Messmer MM, Lee M et al (1991) Diversity and relationships among U.S. Maize inbreds revealed by restriction fragment length polymorphisms. *Crop Sci* 31:669–678
- Messmer MM, Melchinger AE, Herrmann RG, Boppenmaier J (1993) Relationships among early European Maize inbreds: II. Comparison of pedigree and RFLP data. *Crop Sci* 33:944–950
- MMEDV (1999) Molecular and other markers for establishing essential derivation in crop plants (EDV). EU-AgriNet. http://ec.europa.eu/research/agriculture/projects/q1rt_1999_01499_en.htm. Accessed 18 Oct 2012
- Plaschke J, Ganai MW, Roder MA (1995) Detection of genetic diversity in closely related bread wheat using microsatellite markers. *Theor Appl Genet* 91:1001–1007
- Rodrigues DH, de Alcantara Neto F, Schuster I (2008) Identification of essentially derived soybean cultivars using microsatellite markers. *Crop Breed Appl Biotechnol* 8:74–78
- Roldan-Ruiz I, van Eeuwijk FA, Gilliland TJ et al (2001) A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (*Lolium perenne* L.) varieties. *Theor Appl Genet* 103:1138–1150
- Smith JSC, Hussain T, Jones ES et al (2008) Use of doubled haploids in maize breeding: implications for intellectual property protection and genetic diversity in hybrid crops. *Mol Breed* 22:51–59
- Smith JSC, Smith OS (1989) The description and assessment of distances between inbred lines of maize: II. The utility of morphological, biochemical, and genetic descriptors and a scheme for testing distinctiveness between inbred lines. *Maydica* 34:151–161
- Smith JSC, Smith OS (1991) Restriction fragment length polymorphisms can differentiate among U.S. maize hybrids. *Crop Sci* 31:893–899
- Smith JSC, Smith OS, Bowen SL et al (1991) The description and assessment of distances between inbred lines of maize: III. A revised scheme for the testing of distinctiveness between inbred lines utilizing DNA RFLPs. *Maydica* 36:213–226

- Smith OS, Smith JSC, Bowen SL et al (1990) Similarities among a group of elite maize inbreds as measured by pedigree, F1 grain yield, grain yield heterosis, and RFLP's. *Theor Appl Genet* 80:833–840
- Staub JE, Chung S-M, Fazio G (2005) Conformity and genetic relatedness estimation in crop species having a narrow genetic base: the case of cucumber (*Cucumis sativus* L.). *Plant Breed* 124:44–53
- Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277:1063–1066
- Troyer AF, Rocheford TR (2002) Germplasm ownership: Related Corn inbreds. *Crop Sci* 42:3–11
- Tuberosa R, Salvi S (2009) QTL for agronomic traits in Maize production. In: Bennetzen JL, Hake SC (eds) *Handbook of Maize: Its biology*. Springer, New York, pp 179–204
- UPOV (1991) International convention for the protection of new varieties of plants. UPOV. <http://www.upov.int/en/publications/conventions/1991/act1991.htm>. Accessed 22 Oct 2012
- UPOV (2009) Explanatory notes on essentially derived varieties under the 1991 act of the UPOV convention. UPOV/EXN/EDV/1. UPOV, Geneva
- UPOV BMT (2007) EDV in Corn: Concepts of essential derivation and dependence; Possible use of DNA markers: The Maize case. Ad hoc crop subgroup on molecular techniques for maize, Chicago December 3. BMT-TWA/Maize/2/7-a. UPOV, Geneva
- UPOV BMT Add (2007) Guidelines for the handling of a dispute on essential derivation of Maize. Addendum to document BMY-TWA/Maize/2/7/-a. BMT-TWA/Maize/2/7/-a Add, Chicago December 3. UPOV, Geneva
- UPOV BMT (2011) An EDV Court case in wheat in Germany. Presentation by marcel bruins, Secretary general ISF at the UPOV BMT meeting, November 2011
- van Eeuwijk FA, Baril CP (2001) Conceptual and statistical issues related to the use of molecular markers for distinctness and essential derivation. *Acta Hort* 546:35–53
- van Eeuwijk FA, Law JR (2004) Statistical aspects of essential derivation, with illustrations based on lettuce and barley. *Euphytica* 137:129–137
- Wang J, Zhong GY, Chin ECL et al (2002) Identification of parents of F1 hybrids through SSR profiling of maternal and hybrid tissue. *Euphytica* 124:29–34

Chapter 4

Application of Molecular Markers in Spatial Analysis to Optimize *In Situ* Conservation of Plant Genetic Resources

Maarten van Zonneveld, Ian Dawson, Evert Thomas, Xavier Scheldeman, Jacob van Etten, Judy Loo and José I Hormaza

Contents

4.1	Introduction	68
4.2	Application of Molecular Markers to Optimize <i>In Situ</i> Conservation	70
4.3	Geospatial Analysis Techniques for Mapping Molecular Genetic Diversity	72
4.4	Case Study: Climate Change Impact on Cherimoya: Microsatellite Diversity and its Distribution Currently and in the Future	76
4.4.1	Introduction	76
4.4.2	Methods	78
4.4.3	Results and Discussion	80
	References	86

Abstract There is a growing recognition of the need to evaluate the diversity status and trends of plant genetic resources' use and maintenance in natural populations, farmers' fields, home gardens and in other *in situ* settings to prioritize and optimize conservation actions and link these effectively with *ex situ* preservation approaches.

M. van Zonneveld (✉) · J. van Etten
Bioversity International, Turrialba office, Costa Rica
e-mail: m.vanzonneveld@cgiar.org

E. Thomas · X. Scheldeman
Bioversity International, Regional Office for the Americas, Cali, Colombia
Ghent University, Faculty of Bioscience Engineering, Gent, Belgium

I. Dawson
The World Agroforestry Centre, Headquarters, Nairobi, Kenya

J. Loo
Bioversity International, Headquarters, Rome, Italy

J. I. Hormaza
Instituto de Hortofruticultura Subtropical y Mediterránea, (IHSM-UMA-CSIC),
Estación Experimental La Mayora, Algarrobo-Costa, Málaga, Spain

The recent development of new powerful molecular tools that reveal many genome-wide polymorphisms has created novel opportunities for assessing genetic diversity, especially when these markers can be linked to key adaptive traits and are employed in combination with new geo-spatial methods of geographic and environmental analysis. New methods to prioritize varieties, populations and geographic areas for *in situ* conservation, and to enable monitoring of genetic diversity over time and space, are now available to support *in situ* germplasm management of annual crop and tree genetic resources. We will discuss concepts and examples of application of molecular markers and spatial analysis to optimize *in situ* conservation. We present a case study on the distribution and genetic diversity of the underutilized new world fruit tree crop cherimoya (*Annona cherimola* Mill.) in its Andean distribution range to exemplify the usefulness of combining molecular marker and spatial data to inform *in situ* conservation decisions.

Keywords *Annona cherimola* Mill (Cherimoya) · Biogeography · Climate Change · Conservation genetics · Conservation genomics · Domestication · Geographic Information Systems · Germplasm collection · Microsatellite markers · Spatial genetics · Simple sequence repeat

4.1 Introduction

There is an increasing recognition of the need to assess *in situ* diversity status and dynamics of plant genetic resources (PGR) (e.g. in wild populations and on farm) to prioritize and optimize conservation actions and link these effectively with *ex situ* preservation approaches (Palmerberg-Lerche 2008; FAO 2010, 2011). *In situ* PGR are often threatened by the modernization and expansion of agriculture, which involves clearance of more land, replacement of landraces by advanced crop varieties, and new management approaches that exclude diversity from the agricultural landscape, leading to genetic erosion (van de Wouw et al. 2010a).

In situ conservation is considered to be important because it provides dynamism, the potential for continued evolution to natural and human selection pressures. The latter include the requirement for new and better crop varieties and improved trees to meet evolving farmer and market preferences, and anthropogenic climate change (Reed and Frankham 2003; Cleveland and Soleri 2007; Mercer and Perales 2010). This makes *in situ* conservation areas potential sources of untapped and new diversity for the development of new crop varieties for local and wider use.

In situ conservation is also the method of choice for many plants with recalcitrant seed that cannot be stored in seed banks and for plants whose biology (e.g. long period to maturity, large size) otherwise makes human-managed regeneration costly or difficult; these criteria apply to thousands of locally or globally important tree species. In the case of non-timber forest products, genetic resources are often principally maintained in wild stands or, depending on the level of domestication, in

smallholders' fields and home gardens. On farm conservation of tree species within their native distributions is often referred to as *circa situm* rather than *in situ* conservation to make a distinction from preservation in natural populations [Boshier et al. 2004], but in this chapter, we use the term *in situ* to refer to all plant genetic resources - including trees, crops and crop progenitors - in farm and natural settings. The diversity of annual crops and tree species maintained in farms and in the wild is a treasure trove for as yet uncharacterized resources for local people and breeders (Scheldeman et al. 2003; Ræbild et al. 2011). However, trees in modified natural habitats and farmland may be susceptible to particular pressures such as inbreeding depression (Dawson et al. 2009, 2011; Vranckx et al. 2011).

The formulation of *in situ* conservation strategies can be optimized by an understanding of spatial patterns of genetic diversity (Petit et al. 1998). Areas of high genetic diversity may be targets for *in situ* conservation as they may be more likely to contain interesting materials for crop and tree improvement. Measuring genetic diversity *in situ* is also a means for prioritising accessions for *ex situ* collections (Frankel et al. 1995a; Tanksley and McCouch 1997; Odong et al. 2011). Genetic characterization is increasingly being used to optimize *in situ* conservation approaches in combination with new geospatial methods for presenting results (Samuel et al. 2013; Thomas et al. 2012; van Zonneveld et al. 2012). Comparing the genetic diversity that is present *in situ* with what is maintained *ex situ* provides guidance in devising sampling strategies to fill *ex situ* collection gaps (Samuel et al. 2013; van Zonneveld et al. 2012). Similarly, comparisons of farm stands with wild plant populations can demonstrate the relative effectiveness of cultivated and natural landscapes for conservation (e.g. Hollingsworth et al. 2005; Miller and Schaal 2005). At the same time, knowledge on patterns of genetic diversity in the wild and in farmland allows us to better understand the evolutionary processes in the development of current species distributions and, where relevant, in domestication (e.g. Russell et al. 2011). Of course, monitoring activities are also required to measure the effectiveness of *in situ* conservation programs over time, and to account for dynamic processes in the use and management of natural and agricultural landscapes and the transitions between them.

Various initiatives that promote the conservation and sustainable use of plant genetic resources draw attention to the need for more assessments of genetic variation with molecular markers (FAO 2010, 2011). The recent development of new powerful molecular tools that reveal many genome-wide polymorphisms has created novel opportunities for assessing genetic diversity. This is especially the case when these markers can be linked to key adaptive traits and are employed in combination with new geospatial methods of geographic and environmental analysis (e.g. Escudero et al. 2003; Manel et al. 2003; Holderegger et al. 2010; Chan et al. 2011; Tuberosa et al. 2011). New methods to prioritize varieties, populations and geographic areas for *in situ* conservation, and to enable monitoring of genetic diversity over time and space, are now available and can and should be exploited to improve *in situ* germplasm management.

4.2 Application of Molecular Markers to Optimize *In Situ* Conservation

In situ conservation programs should seek to conserve functional genetic variation that is important to foster future adaptive responses in agricultural and natural landscapes and to support human needs such as food security and agricultural productivity in managed ones. Often, though, the variation that will be important in the future is not known currently. As a result, some practitioners have taken the view that simply as much variation as possible, whether of known value or not, should be conserved (e.g. van Zonneveld et al. 2012). In this situation, ‘neutral’ molecular markers, which are not linked to any particular trait but presumably provide a representation of the ‘underlying’ diversity in an organism, are appropriate. Although such markers are not associated with fitness or adaptive potential (Avisé 2010; Ouborg et al. 2010), they contribute to a good understanding of many of the evolutionary processes involved in the development of contemporary patterns of variation, including in the contraction and expansion of populations and the development of refugia. They are also ideal for understanding mating systems, the level of inbreeding and other key biological features of importance for PGR management (Brown and Hodgkin 2008). Such markers also reveal the level of kinship between different crop landraces and the degree of the genetic contribution of different ancestors to cultivars (Eaton et al. 2006). This has for example been used to prioritize livestock breeds for *in situ* conservation on the basis of their genetic distinctiveness (Eding et al. 2002). These methods are now also being applied to crop genetic resources (Samuel et al. 2013).

Allelic richness at neutral loci is often regarded as an indicator of effective population size (Widmer and Lexer 2001; Leberg 2002), which expresses the rate of historic gene flow and bottleneck events. The measure has been used to target wild tree populations for *in situ* conservation (Petit et al. 1998) and to monitor erosion in crop gene pools (van de Wouw et al. 2010a). The number of locally common alleles (alleles that only occur in a limited area of a species’ distribution but reach relatively high frequencies in these areas) has been identified as a particularly useful measure of richness. The maintenance of such alleles at high frequency in particular geographic areas may reflect long processes of selection and adaptation (Frankel et al. 1995b; van de Wouw et al. 2010a). The identification of areas where geographically restricted alleles occur in high frequency can also be determined by Allelic Aggregation Index Analysis (AAIA) (Miller 2005). This calculates for each sampled individual the average proximity of its alleles to similar alleles in other individuals, in comparison to the average distance based on the distribution of all individuals (Miller 2005). When only alleles with a frequency higher than 5% are included in AAIA, the function can be appropriate for calculating locally common alleles.

Although ‘neutral’ markers do not directly relate to function, their heterozygosity can correspond with population fitness, especially for out-breeding species (Reed and Frankham 2003; Vranckx et al. 2011).

At first sight counter-intuitively, increases in morphological variation in key features that are selected by humans in the domestication process of annual crops (e.g. *Brassica*, maize, chilli peppers: *Capsicum* spp., potato: *Solanum tuberosum* L.) and trees (e.g. cacao: *Theobroma cacao* L., apple: *Malus domestica* Borkh.) are often

accompanied by decreases in genetic variation in the wider genome. This apparent paradox has fascinated students of domestication for many years, and it may be due to bottlenecks induced by human transport of germplasm and/or phenotypic selection events. With human selection, the range of variation at traits of interest becomes wider, but elsewhere bottlenecks are introduced (e.g. de Haan et al. 2009a).

Different types of characterization thus provide us different information and insights. Different characterization approaches may be used simultaneously to target areas for *in situ* conservation because each method reveals different features about populations. While some approaches may specifically reveal the results of recent gene flows, other methods may shed light on ancient evolutionary processes that relate to climatic fluctuations over tens or hundreds of thousands of years (Newton et al. 1999). Increasingly, molecular markers are being identified that are linked to genes associated with adaptive traits, which bridges the gap to function. Allelic shifts at loci linked to adaptive traits under selection pressure can be evaluated against changes at neutral reference loci to distinguish ‘real’ adaptive genetic changes from migration and drift, and to separate plastic from genetic responses (Hansen et al. 2012). We return to this topic later.

The use of molecular tools to target areas for *in situ* conservation has a number of practical advantages compared with morphological characterization. First, it is relatively easy to collect the samples needed for molecular analysis in the field and to transport them to the laboratory for testing (easier, e.g. to sample leaves than collect seeds that may be recalcitrant or difficult to germinate). Second, samples can be analysed in a laboratory in another country. Assuming the necessary permissions to exchange material have been obtained. This is particularly useful when examining species’ diversity patterns across extensive distribution ranges covering many countries to ensure consistency in analytical methods. Third, markers are neutral to environmental ‘noise’ that is always present when contrasting the morphological traits of plants grown in different locations. This can lead to plants that look different from each other even when they are genetically very similar. An alternative is to characterize plants in environment-controlled field trials, but these are often expensive and a certain amount of environmental noise may remain. Fourth, modern molecular marker methods are generally repeatable over time and location, which provides opportunities to add data from extra, freshly sampled populations to existing data sets. This is important when monitoring the dynamics of diversity in populations over time, for example, when assessing genetic erosion. The molecular diversity of a ‘historic’ collection from a specific area can be compared with a new one, such as de Haan et al. (2009a) did to assess possible allelic loss over time in local potato varieties grown in Peruvian Andean villages. In this particular case, no loss of molecular diversity was observed over a 25-year period, suggesting *in situ* conservation with farmers was effective. When improved varieties cross with local landraces and are taken up into informal seed systems they may however, reduce *in situ* diversity, as shown for maize in southern Mexico (van Heerwaarden et al. 2009).

Despite examples of reductions of *in situ* crop genetic diversity due to replacement and hybridization with new varieties, levels of newly introduced variation may increase. For example, a meta-analysis of molecular diversity studies of eight food

crops suggested that in the last decades breeders have increased the use of crop diversity in the development of improved varieties (van de Wouw et al. 2010b).

As mentioned in the introduction, on farm conservation can complement the conservation of wild populations increasingly menaced by natural habitat loss. This is relevant for many socio-economically important tree species that are incipient domesticates. Further research is needed on the ecological and socio-economic circumstances under which on farm conservation is an effective approach for the sustainable management of tree genetic resources (Dawson et al. 2013). In the case of the Amazonian tree species *Inga edulis* Mart. (ice-cream bean tree), for example, molecular marker diversity is lower in farms than in wild populations, although allelic variation remains relatively high in agricultural landscapes that are still important sites for conservation (Hollingsworth et al. 2005). In another example, cultivated populations of the Mesoamerican fruit tree *Spondias purpurea* L. (jocote) contained unique chloroplast alleles that were not found in wild populations, supporting the complementary roles of on farm- and wild stand-conservation approaches (Miller and Schaal 2005).

4.3 Geospatial Analysis Techniques for Mapping Molecular Genetic Diversity

Just as molecular marker methods have advanced greatly over the last decade, so too have approaches for geospatial analysis (Guarino et al. 2002; Miller 2005; Jombart 2008; van Etten and Hijmans 2010; Chan et al. 2011; van Zonneveld et al. 2011). Advances in geographic information systems (GIS) are still however underutilised in genetic diversity studies, perhaps because many scientists are unaware of the newer methods available. Training materials have been developed to bridge this gap (Scheldeman and van Zonneveld 2010). A great advantage of GIS-based approaches is the clear graphic presentation of results through maps, which facilitates the interpretation of findings and hence their incorporation into conservation strategies (Jarvis et al. 2010). Geospatial analysis of genetic diversity has been undertaken for a wide range of trees because the maintenance of their genetic resources often depends largely on *in situ* conservation. For the Norway spruce (*Picea abies* [L.] Karst.) in Austria, for example, a geographic grid-based gap analysis was carried out to identify new conservation units that complemented mitochondrial and nuclear molecular marker studies (Schueler et al. 2012).

One effective method to describe genetic diversity in geographic space is to use circular neighbourhood-type analyses. This is especial effective when working with individual geographically-referenced accessions rather than with populations (van Zonneveld et al. 2012). The circular neighbourhood approach allows calculation with confidence of genetic parameters per grid cell by grouping georeferenced individuals within a user-defined radius of geographic distance around each cell (Scheldeman and van Zonneveld 2010). The approach makes analyses less sensitive to grid origin definition and enables the inclusion of isolated trees in the calculation of genetic parameters.

In the approach uneven sampling densities among grid cells can be corrected by establishing a level of Rarefaction (minimum sample size per grid cell to include in analysis) or by carrying out re-sampling without replacement (see Leberg 2002; Thomas et al. 2012; van Zonneveld et al. 2012). The final results of the corrected diversity analysis then provide detailed and representative estimates of geographic patterns of diversity. Scaling can be adjusted to the dimensions of particular countries or regions so that results can be incorporated into national and regional conservation plans, as appropriate. Such an approach has been used to identify genetic diversity hotspots for the *in situ* conservation of a number of important perennial tree crops, including cacao in its Latin-American centres of origin and domestication (Thomas et al. 2012), cherimoya in the Andes (van Zonneveld et al. 2012), and bush mango (*Irvingia gabonensis* [Aubry-Lecomte ex O'Rorke] Baill. and *I. wombolu* Vermoesen) in Central Africa (Lowe et al. 2000).

These are examples of geospatial analyses to prioritize conservation efforts for a few economically important trees. However, thousands of tree species have local livelihood value and many of these are threatened. As the costs to carry out analyses with molecular markers are continuously decreasing, it will become more feasible to perform such studies on these species.

One approach to extrapolate patterns observed from these analyses and prioritize areas with as many tree genetic resources as possible for conservation is to identify putative Pleistocene refugia and converging post-glacial migration routes. These areas may harbour high inter- and intra-specific diversity (Petit et al. 2003). Georeferenced observation points from herbaria and genebanks can be used to predict Pleistocene distributions on the basis of extrapolated past climate data (Waltari et al. 2007). Climate data are freely available from online platforms such as PMIP2 (<http://www.p mip2.cnrs-gif.fr>) and WorldClim (www.worldclim.org). Georeferenced plant data are increasingly available through sites such as the Global Biodiversity Information Facility (www.gbif.org). These data, when of reasonable quality, can be used in Environmental Envelope Modelling (EEM) to predict past distributions and reconstruct potential refugia (Waltari et al. 2007; Thomas et al. 2012; Vinceti et al. 2013). Molecular marker data, especially chloroplast DNA variation, can help to validate or refute potential refugia determined by EEM (Newton et al. 1999; Petit et al. 2003). A major limitation, however, is that different sampling methods and marker types have often been used for separate studies of the same species in different parts of its distribution. This complicates clear identification of distribution-wide diversity patterns, as observed for example for the new world tropical palm *Bactris gasipaes* Kunth (peach palm) (Clement et al. 2010; Graefe et al. 2013). For most important food crops standardized molecular tool kits have been proposed to improve comparability (Van Damme et al. 2010); however, for most other plants, molecular standards still need to be developed.

Most crops were domesticated in the last 12,000 years and the current distribution of their diversity is marked by relatively recent human dispersal. More inter- and intra-specific diversity can be expected to be found in and around centres of domestication such as the Andes, Mesoamerica and the Amazon in the Americas, and the Fertile

Crescent in the Middle East. Just as high tree genetic diversity is expected in post-Pleistocene converging migration routes, so high crop diversity can be expected in converging human dispersal routes.

An example is cultivated chili pepper in Peru. The diversity of cultivated *Capsicum* encountered there is probably the highest in the world. It is an important area of diversification and varieties from the five cultivated species have been grown there since pre-Colombian times (Perry 2012). However, Peru is probably not the centre of origin of these five species. Rather, they were likely transported to their current locations in Peru from putative centres of domestication elsewhere (Eshbaugh 2012). In such situations, molecular markers can help distinguish between centres of origin and centres of diversity.

Studies in human genetics show that relatively simple models of diffusion can be used to predict global genetic diversity patterns (Ramachandran et al. 2005). Diffusion models have been used to model the spread of agriculture generally and of particular crops (Pinhasi et al. 2005). van Etten and Hijmans (2010) showed that, for crops, spatial diffusion models and genetic diversity models can be linked. Such combined models could eventually be used to predict levels of diversity and complementarity between locations, including of un-sampled locations.

Spatial studies with a more local scope can be important for deciding in detail the most appropriate on farm PGR management strategies in traditional rural communities. Such studies can, for example, help to better understand how farmers manage and conserve crop diversity within the landscape over time (Worthington et al. 2012). This can help identify the geographic and social levels at which *in situ* conservation should be implemented and crop diversity monitored (Barry et al. 2007).

PGR management in traditional rural communities differs by crop species, by social context and by environment. For example, molecular markers have shown that farmers in southern Mexico maintain bean diversity (*Phaseolus coccineus* L., *P. dumosus* Macfad., *P. vulgaris* L.) in clearly separated fields along a topographical climate gradient (Worthington et al. 2012). A simple sequence repeat (SSR) analysis of the genetic structure of rice (*Oryza sativa* L., *O. glaberrima* Steud.) in the Republic of Guinea revealed genetic differences between lowland coastal and upland savannah areas, but no differentiation between villages or farms within contrasting agro-ecosystems (Barry et al. 2007). Although within each rice variety high genetic diversity was found, most of this diversity could be conserved within just a few farms of a village. The high diversity within farms and the low genetic structure between farms observed by Barry et al. (2007) may be explained by active human seed exchange and high varietal turnover. Likewise, in potato growing areas in the Peruvian Andes, most variation in SSRs in the potato (*Solanum* spp.) crop was observed within farms (de Haan et al. 2009a). The same study of de Haan et al. (2009a) indicated that the level of diversity maintained by farm families can however vary greatly. The conservation of high potato varietal diversity by many farmers can be explained by preferences for specific cultivars for home consumption and the desire to spread risk through varietal diversification (de Haan et al. 2009a, b).

As some of the above examples illustrate, GIS can be used to overlay different information types on to genetic data to make more informed management decisions.

An understanding of the drivers of genetic erosion of natural populations, threats to ecosystems and their relative vulnerability, can be gained (e.g. Jarvis et al. 2010; Hirota et al. 2011).

Threats to ecosystems need to be interpreted carefully when applied to particular species, since individual taxa, and the populations within them, will respond differently. Nevertheless, areas of important genetic diversity under threat can be identified for urgent conservation, such as natural populations with high allelic richness located in areas of agricultural encroachment and/or in locations where future climate will likely not support regeneration and survival.

Recent studies have begun to explore further how to incorporate spatially-defined threat information to prioritize for *in situ* conservation. Optimal solutions for conservation considering the costs to conserve unique genetic diversity can be calculated (Samuel et al. 2013). Information on the probability of variety replacement can be included, based on variables such as the geographic distance to areas of high human population density (Samuel et al. 2013).

EEM of species distributions within current and projected future climates to assess changes over time can be used in combination with genetic analysis to identify hotspots of diversity that are particularly vulnerable to change. This has for example been done for cacao (Thomas et al. 2012) and a further example (cherimoya) is described below. The comparison of current and future modelled distributions for cacao revealed several areas of low climate change threat within the Amazonian area with high genetic diversity; these areas could be targets for *in situ* conservation (Thomas et al. 2012). Tree species are good candidates for studying climate change impacts on landscapes because of their longevity, which means signatures of past events are retained for longer (Petit et al. 2008). For trees, the available molecular data in combination with pollen cores and other data sets would suggest that natural dispersal will not be able to keep up with climate change in many parts of the world. Whole forest ecosystems that are crucial for the *in situ* conservation of trees and associated flora (including the wild populations of some crucial crops and their relatives) and fauna may therefore be threatened (Malhi et al. 2009).

For annual crops, a good example of application of current and Past climate EEM is with wild barley (*Hordeum vulgare* L. subsp. *spontaneum* [K. Koch] Thell.) in the Fertile Crescent and Central Asia (Russell et al. 2013). In this case, contemporary patterns of molecular marker diversity expressed using the circular neighbourhood method corresponded with EEM for the Last Glacial Maximum. Both analyses indicated that the eastern Mediterranean was likely to have been a Pleistocene refuge for wild barley, as the highest levels of genetic diversity were located here and habitat was indicated to be common. This area should therefore be a focus for conservation activities.

Most interestingly in this case, geographic point location data of wild barley accessions were used to identify the environments in which the taxon grows in its by extracting values for the 19 bioclimatic variables of WorldClim. The advantage of studying barley compared to many lesser-researched plants is that the chromosome positions of many molecular markers are known. This allows associations between environmental data and genetic markers to be located within the genome. This has

the potential to be very useful in crop breeding and in monitoring responses to environmental change (Hansen et al. 2012). In the case of wild barley, for example, it was possible to identify regions of the genome that are candidates for adaptive genes to the environment. This type of analysis may prove especially useful for plants for which comparatively little phenotypic data are available (Neale and Kremer 2011). Phenotypic data taken directly from wild stands rather than collected from field trials will also become more important.

The study of genetic and plastic responses of plant populations to climate change is especially relevant when migration to more suitable locations may be restricted due to habitat fragmentation and/or by the rapid pace of change (Hoffmann and Sgró 2011). Distribution range shifts may also cause reduced fitness of populations due to founder effects (Cobben et al. 2011). Molecular data modelled in geographic space can help determine potential migration rates and adaptation at current locations can be monitored through allele shifts at important genes (as described in the barley example above). The latter approach is becoming increasingly feasible as chromosome-mapped markers are linked to adaptive traits.

Conservation genomics (i.e. combining conservation genetic principles with functional genomics approaches) is in Ouborg's (2010) opinion both necessary and feasible to understand the effects of genetic diversity losses on fitness. Avise (2010) noted that the 'genomics revolution' allows scientists to examine sequence variation at unprecedented numbers of loci for unprecedented numbers of individuals. Although most genomic advances are currently associated with well-studied crops and some other model species, rapid developments will allow for genome-wide mapping in most plant species in the near future (Ingvarsson and Street 2011).

A major challenge for mapping diversity at gene sequences of adaptive significance is that important traits may be controlled by many loci. If drought tolerance is influenced by more than 200 loci, for example, what is the value of choosing only a handful of these, to study? Such concerns will be less relevant with the application of 'exome capture' methods that assess variation among individuals at all expressed genes and with new statistical approaches to assess genome-environment associations (Mascher et al. 2013).

4.4 Case Study: Climate Change Impact on Cherimoya: Microsatellite Diversity and its Distribution Currently and in the Future

4.4.1 Introduction

In this section we present a case study on the distribution and genetic diversity of cherimoya in its Andean range to exemplify the usefulness of combining molecular marker and spatial data to inform *in situ* conservation decisions. Northern Peru (Cajamarca) and southern Ecuador (Loja) were identified as areas of high conservation priority for cherimoya and as important areas for further germplasm exploration on the basis that areas of high neutral genetic diversity have a high likelihood of

containing unknown traits of interest for domestication (van Zonneveld et al. 2012). Here, we compare a climate change impact study of cherimoya with the spatial distribution of its genetic diversity in the Andes and discuss the potential implications for conservation. Temperature and rainfall variations are likely to occur in high genetic diversity areas of cherimoya's distribution due to climate change, potentially threatening genetic resources.

Cherimoya is an underutilized new world tropical fruit tree belonging to the Annonaceae, a family included within the Magnoliales in the Eumagnoliid clade among the early-divergent angiosperms (Bremer et al. 2009). It is still in the initial stages of domestication (Escribano et al. 2007) and is considered at high risk of genetic erosion (Popenoe et al. 1989). Cherimoya fruits are widely praised for their excellent organoleptic characteristics, and the species is therefore considered to have high potential for commercial production and income generation for both small and large-scale growers in subtropical climates (Van Damme and Scheldeman 1999). Cherimoya presents protogynous dichogamy, i.e. it has hermaphroditic flowers wherein female and male parts do not mature simultaneously, favouring outcrossing in its native range (Lora et al. 2010). For commercial production outside of the tree's native range, hand pollination with pollen and stamens is common practice due to a lack in overlap of female and male stages and the absence of pollinating agents (Lora et al. 2010). At present, large-scale commercial production is concentrated in Spain, the world's largest cherimoya producer with around 3000 ha of plantations, while small-scale cultivation occurs throughout the Andes, Central America and Mexico. Cherimoya is commonly grown in Andean home gardens and orchards, and trees from these environments may contain promising traits for future breeding programs (Scheldeman et al. 2003). In Peru, the local cultivar 'Cumbe' is already sold for retail prices significantly above the prices of unselected cherimoya fruit types (Vanhove and Van Damme 2009, 2013).

Most early chroniclers and scientists have proposed the Andean region, more precisely the valleys of southern Ecuador and northern Peru, as cherimoya's centre of origin (Popenoe 1921; Popenoe et al. 1989). The existence of isolated putatively wild cherimoya forest patches in the inter-Andean valleys of Ecuador and northern Peru supports this hypothesis. Nonetheless, the possibility that these are feral populations cannot be immediately discounted, because this phenomenon has been observed for several fruit trees including olives (Gepts 2003). An alternative hypothesis for the centre of origin of cherimoya is Central America, considering that most relatives of cherimoya are native to that region and southern Mexico (H. Rainer, Institute of Botany, University of Vienna, 2011, pers. comm.), and that high genetic diversity is found in cherimoya genotypes there (Hormaza et al., unpublished data). In any case, cherimoya fruits have been consumed in the Andean region since antiquity (Popenoe et al. 1989) and movement of germplasm across southern Mexico, Mesoamerica and the Andes probably took place in pre-Columbian times. Wolters (1999) indicated that the ceramic cherimoya-shaped vases found in the remains to the Ecuadorian Valdivia culture (3,500–1,600 A.C.) may testify to the important role this early culture played in the exchange of cherimoya germplasm (as well as other crops) between the Andean region and Mesoamerica.

In contrast to most tropical and subtropical underutilized fruit trees, cherimoya genetic resources are well represented in *ex situ* germplasm collections. Several field collections have been established in Spain, Peru and Ecuador, preserving over 500 different accessions (Escribano et al. 2007; CHERLA 2008). The Spanish collection, based at la Estación Experimental La Mayora in Malaga, holds over 300 accessions (190 collected from the Andean region) and is currently used as source material for the Spanish cherimoya breeding program. This collection has been thoroughly analyzed using isozymes (Pascual et al. 1993; Perfectti and Pascual 1998, 2005) and Microsatellite markers (Escribano et al. 2007, 2008).

4.4.2 Methods

4.4.2.1 Sampling and SSR Analysis

A total of 1,506 cherimoya accessions were sampled (395 from Bolivia, 351 from Ecuador and 760 from Peru) and tested with nine highly polymorphic nuclear SSR markers (Escribano et al. 2008). DNA was extracted from young leaves according to Viruel et al. (2004). Details of the methods of SSR selection, amplification and PCR product analysis are given in van Zonneveld et al. (2012). The coordinates of tree locations were verified with DIVA-GIS (www.diva-gis.org) discarding erroneous points, i.e., (1) points showing inconsistencies between the location mentioned in passport data and map projections at department and province level, applying a buffer of 20 minutes (approximately 30 km); and (2) outliers based on current climate data derived from WorldClim (Hijmans et al. 2005) (for two or more of the 19 bioclimatic variables according to the reverse jackknife method implemented in DIVA-GIS; Chapman 2005). Based on these checks, only two points were excluded from further analysis. The cleaned dataset thus included microsatellite data of 1,504 georeferenced trees.

4.4.2.2 Spatial Analysis

Similar to van Zonneveld et al. (2012), we constructed 10-minute grid layers (which corresponds to approximately 18 km in the study area) for all genetic parameters, applying a circular neighbourhood with a diameter of one degree (corresponding to approximately 111 km), using R program version 2.15.1 with the packages Raster (Hijmans and van Etten 2012) and Adegenet (Jombart et al. 2013). We performed grid cell-based calculations of allelic richness and the number of locally common alleles as measures of *alpha* genetic diversity. To establish comparability of these parameters between cells, we corrected sample-bias through re-sampling without replacement after Leberg (2002) to a sample size of 20 trees. Per parameter, we calculated the average value from 1,000 subsamples following the bootstrap method developed by

Thomas et al. (2012). Re-sampling without replacement provides similar results to the rarefaction approach applied by van Zonneveld et al. (2012) with the advantage that it can be used to correct other genetic parameters in addition to allelic richness (Thomas et al. 2012). As a measure of *beta* diversity, a spatial principal component analysis (sPCA) was carried out with Adegenet using the neighbourhood-by-distance connection network as explained by Jombart (2013) with a minimum distance of zero and maximum distance of one degree. For each tree the projection score on the first axis was visualized with a 10-minute resolution raster applying a circular neighbourhood of one degree.

We carried out EEM to assess potential impacts of climate change across cherimoya's Andean range by comparing distributions under current and future climate. We used the EEM software Maxent version 3.3.3k (Phillips et al. 2006; Elith et al. 2011) implemented in the R package Dismo (Hijmans et al. 2013), on the basis that it performs very well in comparison to other EEM software (Elith et al. 2006; Aguirre-Gutiérrez et al. 2013). We trained our model on the basis of the 1,504 retained georeferenced cherimoya trees and the 19 WorldClim bioclimatic variables at a resolution of 2.5 minutes (Hijmans et al. 2005). For characterising future climate we used 19 downscaled climate models for 2050 based on the A2 scenario of greenhouse gas emissions (available at <http://ccafs-climate.org>). To improve model performance (cf. Acevedo et al. 2012), we limited the extraction of background points to the area enclosed by the convex hull polygon constructed based on all records and extended with a buffer corresponding to 10 % of the polygon's longest axis (Willis et al. 2003). Apart from this pre-selection of background points, we used Maxent default settings.

To compare unbiased cAUC values and hence the performance of models constructed with Maxent with a geographical null model (see Hijmans 2012), we (1) randomly partitioned both presence and background points in five groups, (2) removed spatial sorting bias; and (3) ran both models for each of the five data subsets (each time using 80% of the points as test data and 20% as training data) using relevant functions implemented in the Dismo package. The Maxent models performed significantly better (mean cAUC = 0.70) than the geographical null models (mean cAUC = 0.51) (Mann-Whitney $P = 0.01$), justifying the use of Maxent. We projected the Maxent model both to (1) the average of the 19 downscaled general circulation climate models for 2050; and (2) to each of the 19 models separately. The modelled distribution based on (1) was slightly more conservative than the distribution obtained by cell-based averaging of the logistic values of (2), and we therefore used (1) for comparison with the present-day modelled distribution. At the same time, the results for each separately run general circulation climate model were used to examine the level of agreement between models in predicting suitable areas for cherimoya in the 2050s. We restricted the potential distributions generated by Maxent using the maximum of the sum of sensitivity and specificity as a threshold value (here 0.15 for the logistic threshold). To reduce the risk of including areas within the modelled distribution where the species does not in reality occur (e.g. due to dispersal limitations), we limited the potential distributional under current climatic conditions to the area enclosed by the convex hull polygon created on the basis of the species' presence

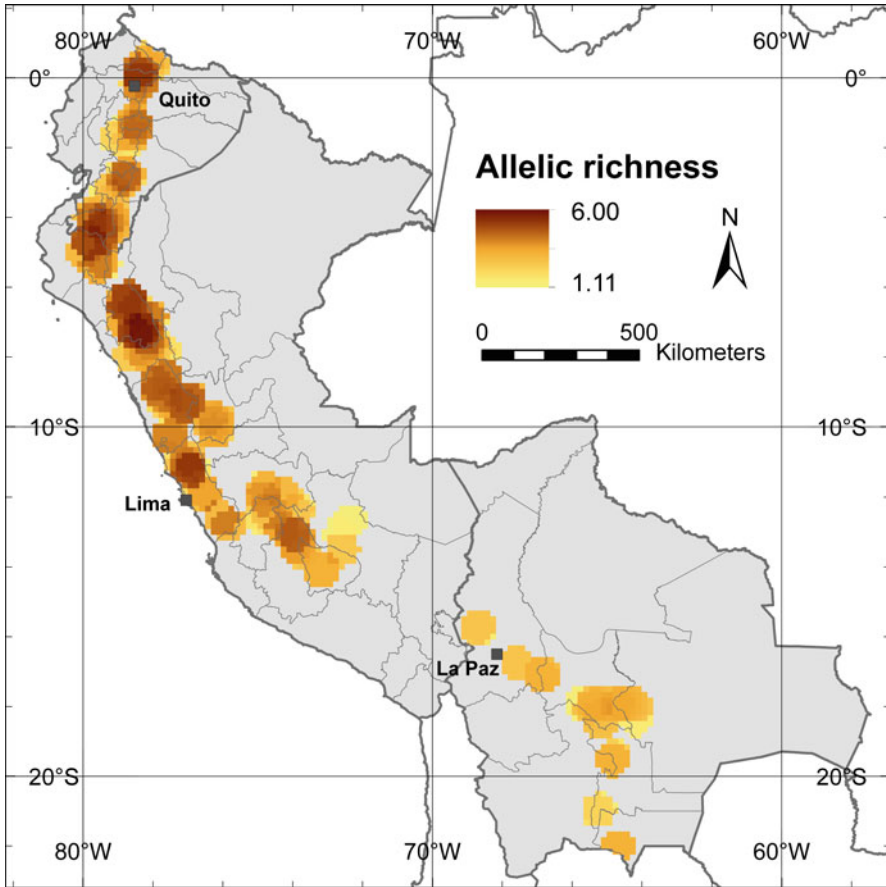


Fig. 4.1 This map shows the average number of alleles per locus in 10-minute grid cells applying a one-degree circular neighbourhood

points and extended with a buffer around it corresponding to 10 % of the polygon's longest axis.

All maps were edited in ArcMap.

4.4.3 Results and Discussion

The cherimoya populations in northern Peru, around the Cajamarca Department, contain the highest allelic richness across the Andean distribution range (Fig. 4.1). Other areas of high Alpha diversity are located on the border zone between Ecuador (Loja Province) and Peru (the Piura Department), in the northern part of Ecuador around the capital Quito and in the northern part of the Lima Department in Peru. However,

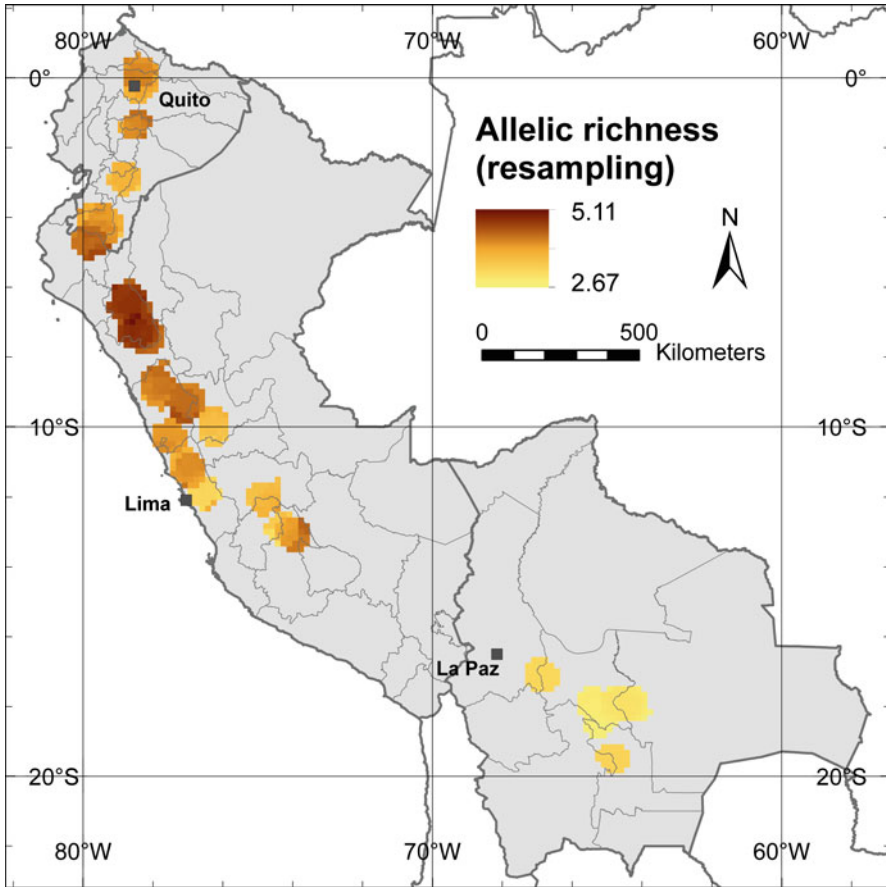


Fig. 4.2 This map shows the average number of alleles per locus in 10-minute grid cells applying a one-degree circular neighbourhood and resampling without replacement to a minimum sample size of 20 trees. The value per grid cell is the average of 1,000 bootstrapped subsamples

when we observe allelic richness corrected by re-sampling without replacement, clearly most diversity is found in northern Peru, around the Cajamarca Department (Fig. 4.2). This parameter is highly correlated with allelic richness corrected by rarefaction to a minimum sample size of 20 trees, as calculated by comparing rarefaction and bootstrapped subsample values for 654 grid cells (van Zonneveld et al. 2012; $r = 0.98$, $P < 0.0001$). These results are confirmed by the high number of Locally common alleles after resampling without replacement found in the Cajamarca department (Fig. 4.3), indicating that the tree stands in this region have likely been subjected to long processes of selection and local adaptation, resulting in the accumulation of diversity.

Beta diversity explained by projection scores on the first axis of the sPCA reveals a clear global genetic structure and little local structure (Fig. 4.4), distinguishing

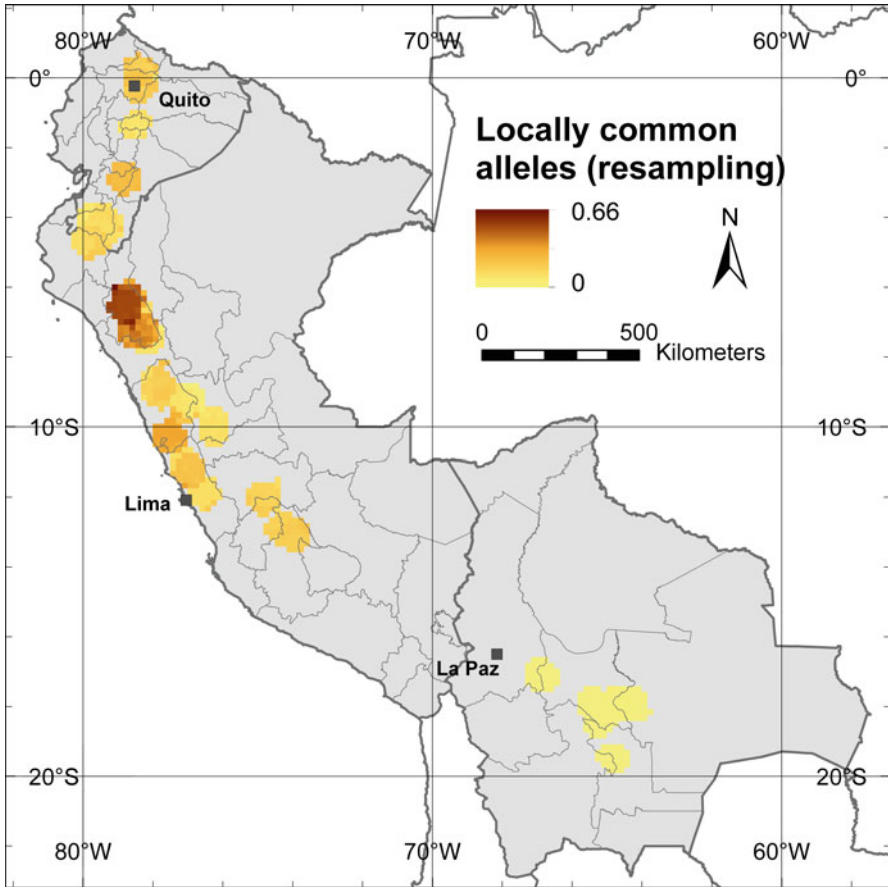
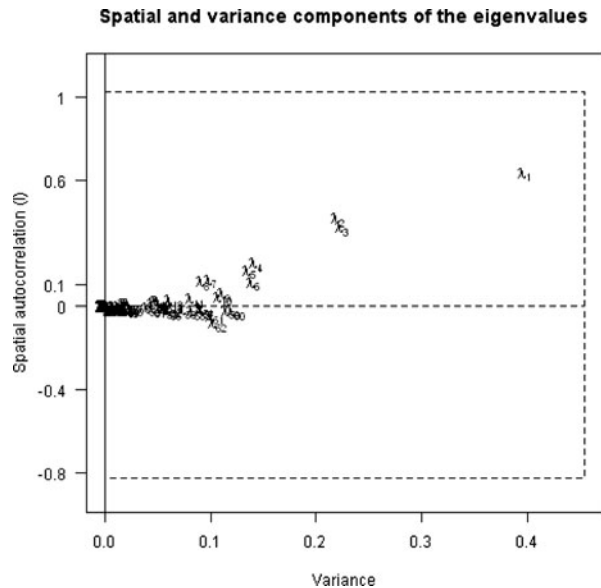


Fig. 4.3 This map shows the average number of alleles per locus in 10-minute grid cells that are relatively common (occurring with a frequency higher than 5% in a limited area (in 25% or less of grid cells), applying a one-degree circular neighbourhood and resampling without replacement to a minimum sample size of 20 trees. The value per grid cell is the average of 1,000 bootstrapped subsamples

between two genetic populations, one comprising northern Peru and Ecuador and the other consisting of southern Peru and Bolivia (Fig. 4.5). This is consistent with the genetic structure found by van Zonneveld et al. (2012) applying Bayesian cluster analysis. The proportion of variance explained by the first sPCA axis by respectively spatial autocorrelation and genetic variance is 0.64 and 0.40 (Fig. 4.4). This indicates clear spatial and genetic structure. The low genetic diversity in Bolivia compared to Peru (and to a lesser degree compared to Ecuador), suggests that populations in Bolivia have been established more recently. Our results from the sPCA suggest that plant material in Bolivia most likely has been introduced from southern Peru,

Fig. 4.4 This graph shows the Eigenvalues (λ) of the spatial principle component analysis for each component according to the genetic variance explained (x -axis) and Moran's Index for spatial autocorrelation (I) (y -axis). Eigenvalues that explain global spatial structure have a positive I, Eigenvalues that contribute to local spatial structure have a negative I



particularly from the regions of Cuzco, Huancavelica and/or Junín, because trees from these Peruvian departments are genetically closely related to Bolivian stands.

Modelling of expected climate change impacts on the cherimoya distribution by the 2050s reveals that with little of its Andean range is expected to be seriously affected, with the exception of a few lower-altitude zones (Fig. 4.6). This suggests that climate change is not a significant threat to cherimoya genetic resources in the region in the next four decades, including in the diversity hotspot in northern Peru. Of course, this analysis includes a certain amount of imprecision because climate prediction models may not capture all local dynamics in inter-Andean valleys and the A2 scenario may be exceeded. Progressive climate change is, of course, expected to continue after the 2050s, and modelling of long-term climate change impacts on cherimoya tree stands is required. Temperatures above 30 °C usually result in pollination problems and can cause the drop of recently set fruit; humidity changes also influence the reproductive process (Lora et al. 2009, 2011, 2012).

Other threats such as replacement by currently more profitable crops such as avocado (*Persea americana* Mill.) may be more important drivers of cherimoya genetic erosion than climate change (personal observation of X. Scheldeman). Possible *in situ* conservation interventions to reduce genetic erosion include high-value market development for traditional cultivars and the organization of seed fairs to promote seed exchange among farmers (for further discussion see van Zonneveld et al. 2012). Several new geographic areas are predicted to have a suitable climate for cherimoya in the 2050s, such as in the high Andes around Lake Titicaca (Fig. 4.7). The future expansion of suitable habitat in the Andes shows the potential for increasing cherimoya cultivation in the region. This could be a good alternative for commercial

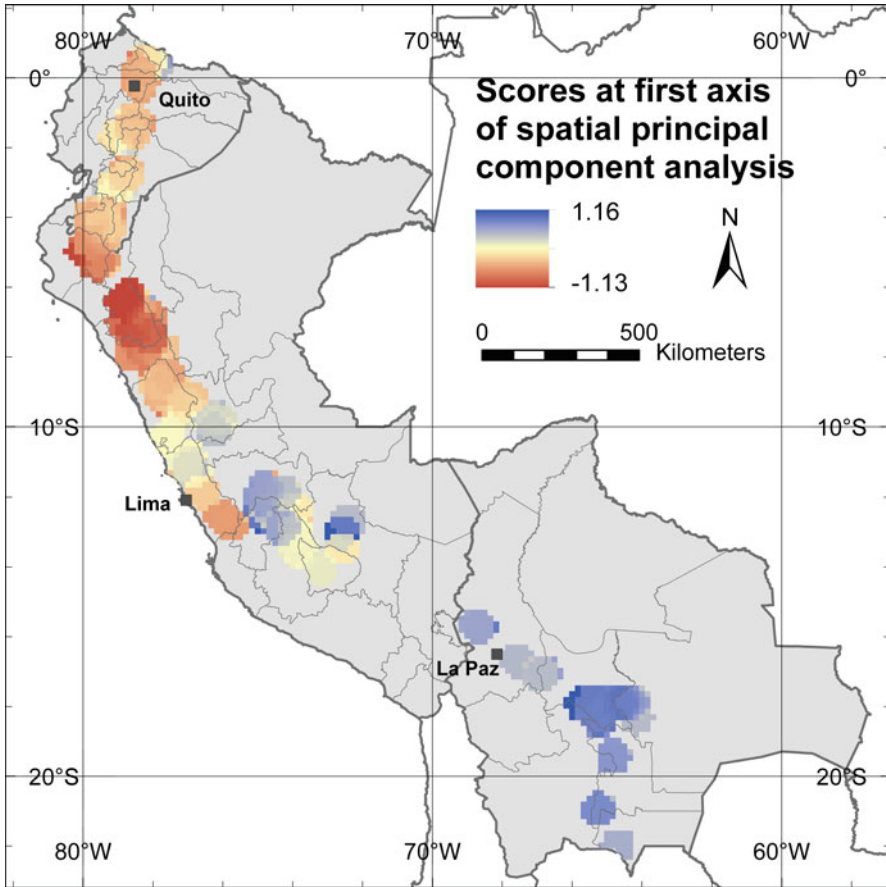


Fig. 4.5 This map shows the average Eigenvector score of trees on the first axis of the spatial principal component analysis for each 10-minute grid cell with 20 or more trees, applying a one-degree circular neighbourhood

cherimoya cultivation, to the Mediterranean countries where commercial production is now centered, but where climate may become too warm and dry in the future.

Our case study shows that hotspots of genetic diversity can be clearly identified with the use of spatial analysis tools, and threats to diversity can be assessed when such analysis is combined with other types of geographic information. In our example, we assessed the impact of climate change on cherimoya's spatial genetic diversity pattern in the Andes. Our results suggest that for cherimoya in its Andean distribution range, climate change impacts may be positive because of an extension of habitat (and reflecting the wide habitat range of the species). Several high-elevation areas are, for example, expected to become newly climatically suitable for cherimoya cultivation in the 2050s. Cherimoya cultivation under a shifting climate may however require the realignment of cherimoya ecotypes adapted to specific climates.

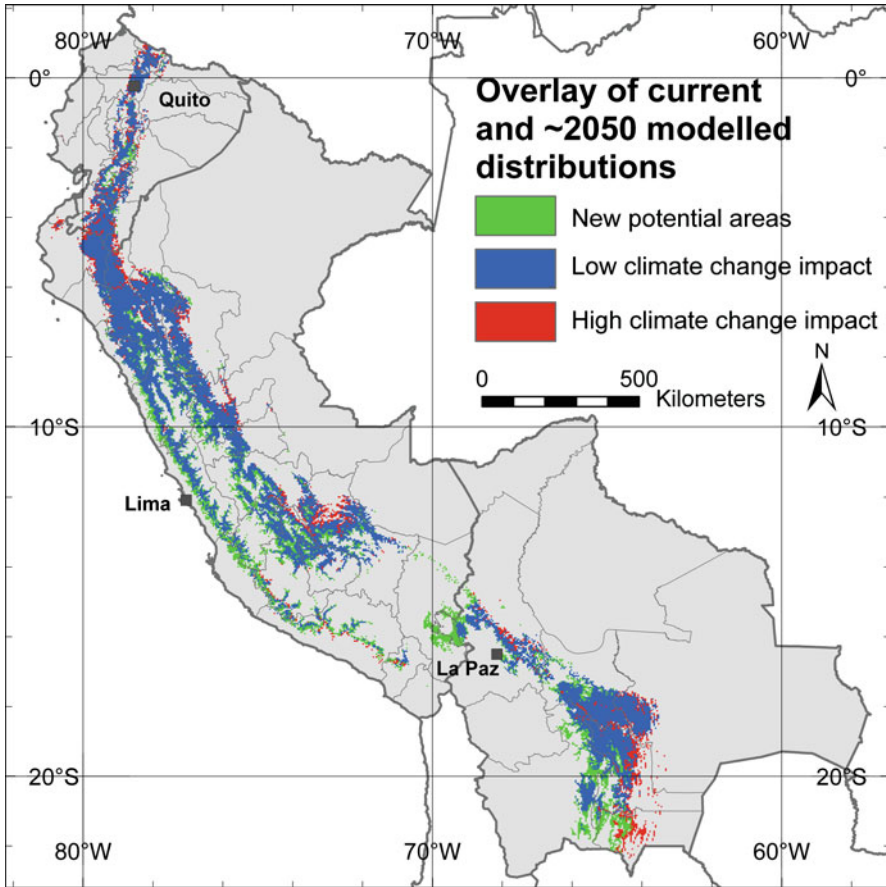


Fig. 4.6 This map provides an overlay of the modelled distributions of cherimoya under current and future (2050s) climatic conditions based on an average of 19 general circulation climate models and emission scenario A2 for future modelling. The map shows where new potential habitat is expected, which current-habitat are expected to remain climatically suitable (low impact), and which current areas are expected to become climatically unsuitable (high impact)

Acknowledgements We thank Patrick Van Damme for his comments on an early version of this chapter. Maarten van Zonneveld thanks the CGIAR research programs Forest, Trees and Agroforestry (FTA) and Climate Change for Agriculture and Food Security (CCAFS) for financial support.

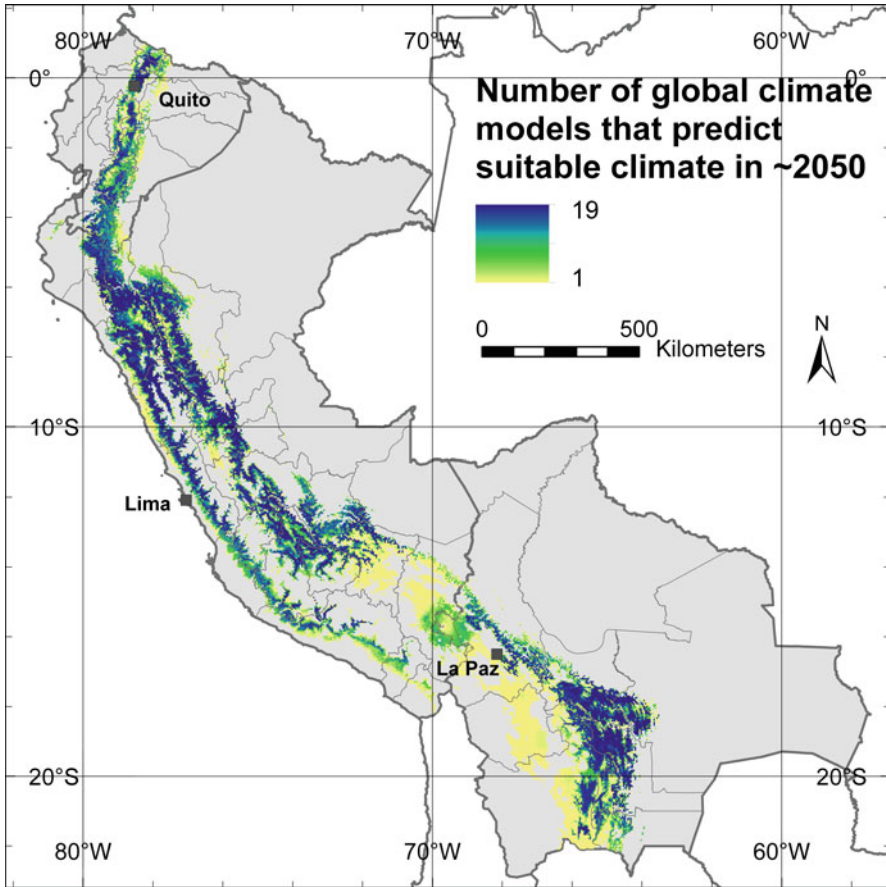


Fig. 4.7 This map shows the number of general circulation global climate models predicting suitable cherimoya habitat in the 2050s under emission scenario A2, based on separate modelling with Maxent for each of the 19 models. The higher the number of models that agree on suitable climate in the 2050s, the more confident the predictions

References

- Acevedo P, Jiménez-Valverde A, Lobo JM, Real R (2012) Delimiting the geographical background in species distribution modelling. *J Biogeogr* 39:1383–1390
- Aguirre-Gutiérrez J, Carvalheiro LG, Polce C et al (2013) Fit-for purpose: species distribution model performance depends on evaluation criteria – Dutch hoverflies as a case study. *PLoS ONE* 8: e63708
- Aguirre-Gutiérrez J, Carvalheiro LG, Polce C et al (2013) Fit-for purpose: species distribution model performance depends on evaluation criteria—Dutch hoverflies as a case study. *PLoS ONE* 8:e63708
- Avise JC (2010) Perspective: conservation genetics enters the genomics era. *Conserv Genet* 11:665–669

- Barry MB, Pham JL, Courtois B et al (2007) Rice genetic diversity at farm and village levels and genetic structure of local varieties reveal need for *in situ* conservation. *Genet Resour Crop Ev* 54:1675–1690
- Bremer B, Bremer K, Chase MW et al (2009) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. *Bot J Linn Soc* 161:105–121
- Boshier DH, Gordon JE, Barrance AJ (2004) Prospects for *circa situm* tree conservation in Mesoamerican dry-forest agro-ecosystems. In: Frankie GW, Mata A, Vinson SB (eds) *Biodiversity conservation in Costa Rica*. University of California Press, Berkeley, pp 210–226
- Brown AHD, Hodgkin T (2008) Measuring, managing and maintaining crop genetic diversity on farm. In: Jarvis DI, Padoch C, Cooper HD (eds) *Managing biodiversity in agricultural ecosystems*, Columbia University Press, pp 13–33
- Chan LM, Brown JL, Yoder AD (2011) Integrating statistical genetic and geospatial methods brings new power to phylogeography. *Mol Phylogenet Evol* 59:523–537
- Chapman AD (2005) Principles and methods of data cleaning—primary species and species-occurrence data, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen
- CHERLA (2008) Inventory of current ex situ germplasm collections. Deliverable 7, Project no. 015100, INCO sixth framework programme
- Clement CR, De Cristo-Araújo M, Coppens D’Eeckenbrugge G et al (2010) Origin and domestication of native Amazonian crops. *Diversity* 2010 2:72–106
- Cleveland DA, Soleri D (2007) Extending Darwin’s analogy: bridging differences in concepts of selection between farmers, biologists, and plant breeders. *Econ Bot* 61:121–136
- Cobben MMP, Verboom J, Opdam PFM et al (2011) Projected climate change causes loss and redistribution of genetic diversity in a model metapopulation of a medium-good disperser. *Ecography* 34:920–932
- Dawson IK, Lengkeek A, Weber JC, Jamnadas R (2009) Managing genetic variation in tropical trees: linking knowledge with action in agroforestry ecosystems for improved conservation and enhanced livelihoods. *Biodivers Conserv* 18:969–986
- Dawson IK, Vinceti B, Weber JC et al (2011) Climate change and tree genetic resource management: maintaining and enhancing the productivity and value of smallholder tropical agroforestry landscapes. A review. *Agroforest Syst* 81:67–78
- Dawson IK, Guariguata MR, Loo J et al (2013) What is the relevance of smallholders’ agroforestry systems for conserving tropical tree species and genetic diversity in *circa situm*, *in situ* and *ex situ* settings? A review. *Biodivers Conserv* 22:301–324
- de Haan S, Núez J, Bonierbale M, Ghislain M (2009a) Species, morphological and molecular diversity of Andean potatoes in Huancavelica, central Peru. In: de Haan S (ed) *Potato diversity at height: multiple dimensions of farmer-driven in-situ conservation in the Andes*, PhD thesis, Wageningen University, The Netherlands, pp 35–58
- de Haan S, Bonierbale M, Juárez H et al (2009b) Annual spatial management of potato diversity in Peru’s central Andes. In: de Haan S (ed) *Potato diversity at height: multiple dimensions of farmer-driven in-situ conservation in the Andes*, PhD thesis, Wageningen University, The Netherlands, pp 91–115
- Eaton D, Windig J, Hiemstra SJ et al (2006) Indicators for livestock and crop biodiversity centre for genetic resources, CGN report 2006/05. Centre for Genetic Resources, CGN/DLO Foundation, Wageningen, The Netherlands
- Eding H, Crooijmans R (2002) Assessing the contribution of breeds to genetic diversity in conservation schemes. *Genet Sel Evol* 34:613–633
- Elith J, Phillips SJ, Hastie T et al (2011) A statistical explanation of MaxEnt for ecologists. *Divers Distrib* 17:43–57
- Elith J, Graham CH, Anderson RP et al (2006) Novel methods improve prediction of species’ distributions from occurrence data. *Ecography* 29:129–151
- Escribano P, Viruel MA, Hormaza JI (2007) Molecular analysis of genetic diversity and geographic origin within an ex situ germplasm collection of cherimoya by using SSRs. *J Am Soc Hortic Sci* 132:357–367

- Escribano P, Viruel MA, Hormaza JI (2008) Development of 52 new polymorphic SSR markers from cherimoya (; Mill.). Transferability to related taxa and selection of a reduced set for DNA fingerprinting and diversity studies. *Mol Ecol Resour* 8:317–321
- Escudero A, Iriondo JM, Torres ME (2003) Spatial analysis of genetic diversity as a tool for plant conservation. *Biol Conserv* 113:351–365
- Eshbaugh WH (2012) The taxonomy of the genus *Capsicum*. In: Russo VM (ed) *Peppers, production and uses*. CABI, pp 1–13
- FAO (2010) The second report on the state of the world's plant genetic resources for food and agriculture. Rome
- FAO (2011) Draft updated global plan of action for the conservation and sustainable utilization of plant genetic resources for food and agriculture. Fifth session of the Intergovernmental Technical Working Group on Plant Genetic Resources for Food and Agriculture, Rome, 27–29 April 2011
- Frankel OH, Brown AHD, Burdon J (1995a) The conservation of cultivated plants. In: Frankel OH, Brown AHD, Burdon J (eds) *The conservation of plant biodiversity*, 1st edn. Cambridge University Press, UK, pp 79–117
- Frankel OH, Brown AHD, Burdon J (1995b) The genetic diversity of wild plants. In: Frankel OH, Brown AHD, Burdon J (eds) *The conservation of plant biodiversity*, 1st edn. Cambridge University Press, UK, pp 10–38
- Gepts P (2003) Crop domestication as a long-term selection experiment. In: Janick J (ed) *Plant breeding reviews 24 Part 2: Long-term selection: crops, animals, and bacteria*, pp 1–44
- Graefe S, Dufour D, van Zonneveld M. (2013) Peach palm (*Bactris gasipaes*) in tropical Latin America: implications for biodiversity conservation, natural resource management and human nutrition. *Biodivers Conserv*. doi:10.1007/s10531-012-0402-3
- Guarino L, Jarvis A, Hijmans RJ, Maxted N (2002) Geographic information systems (GIS) and the conservation and use of plant genetic resources. In: Engels JMM, Ramanatha RV, Brown AHD, Jackson MT (eds) *Managing plant genetic diversity*. International Plant Genetic Resources Institute (IPGRI), Rome, pp 387–404
- Hansen MM, Olivieri I, Waller DM et al (2012) Monitoring adaptive genetic responses to environmental change. *Mol Ecol* 21:1311–1329
- Hijmans RJ, Cameron SE, Parra JL et al (2005) Very high resolution interpolated climate surfaces for global land areas. *Int J Climatol* 25:1965–1978
- Hijmans RJ (2012) Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. *Ecology* 93:679–688
- Hijmans RJ, van Etten J (2012) Geographic analysis and modeling with raster data. R package “Raster.” (<http://cran.r-project.org/web/packages/raster/raster.pdf>)
- Hijmans RJ, Phillips S, Leathwick J, Elith J (2013) Species distribution modelling with R. R package “Dismo.” (<http://cran.r-project.org/web/packages/dismo/dismo.pdf>)
- Hirota M, Holmgren M, van Nes EH, Scheffer M (2011) Global resilience of tropical forest and savanna to critical transitions. *Science* 334:232–235
- Hoffmann AA, Sgró CN (2011) Climate change and evolutionary adaptation. *Nature* 479:479–485
- Holderegger R, Buehler D, Gugerli F, Manel S (2010) Landscape genetics of plants. *Trends Plant Sci* 15:675–683
- Hollingsworth PM, Dawson IK, Goodall-Copestake WP et al (2005) Do farmers reduce genetic diversity when they domesticate tropical trees? A case study from Amazonia. *Mol Ecol* 14:497–501
- Ingvarsson PK, Street NR (2011) Association genetics of complex traits in plants. *New Phytol* 189:909–922
- Jarvis A, Touval JL, Castro SM (2010) Assessment of threats to ecosystems in South America. *J Nat Conserv* 18:180–188
- Jombart T (2008) Adegnet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405
- Jombart T (2013) A tutorial for the spatial analysis of principal components (sPCA) using *adegenet* 1.3–6. R vignette. <http://cran.r-project.org/web/packages/adegenet/vignettes/adegenet-sPCA.pdf>

- Jombart T, Ahmed I, Cori A. (2013) Adegenet: an R package for the exploratory analysis of genetic and genomic data. R package “Adegenet.” <http://cran.r-project.org/web/packages/adegenet/adegenet.pdf>
- Leberg PL (2002) Estimating allelic richness: Effects of sample size and bottlenecks. *Mol Ecol* 11:2445–2449
- Lora J, Herrero M, Hormaza JI (2009) The coexistence of bicellular and tricellular pollen in *Annona cherimola* Mill. (Annonaceae): Implications for pollen evolution. *Am J Bot* 96:802–808
- Lora J, Hormaza JI, Herrero M (2010) The progamic phase of an early-divergent angiosperm, *Annona cherimola* (Annonaceae). *Ann Bot* 105:221–231
- Lora J, Herrero M, Hormaza JI (2011) Stigmatic receptivity in a dichogamous early-divergent angiosperm species, *Annona cherimola* Mill. (Annonaceae). Influence of temperature and humidity. *Am J Bot* 98:265–274
- Lora J, Herrero M, Hormaza JI (2012) Pollen performance, cell number, and physiological state in the early-divergent angiosperm *Annona cherimola* Mill. (Annonaceae) are related to environmental conditions during the final stages of pollen development. *Sex Plant Reprod* 25:157–167
- Lowe AJ, Gillies ACM, Wilson J, Dawson IK (2000) Conservation genetics of bush mango from central/west Africa: implications from random amplified polymorphic DNA analysis. *Mol Ecol* 9:831–841
- Malhi Y, Aragao LEOC, Galbraith D et al (2009) Exploring the likelihood and mechanism of a climate-change-induced dieback of the Amazon rainforest. *Proc Natl Acad Sci U S A* 106:20610–20615
- Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol Evol* 18:189–197
- Mascher M, Richmond TA, Gerhardt DJ et al (2013) Barley whole exome capture: a tool for genomic research in the genus *Hordeum* and beyond. *Plant J* (in press)
- Mercer KL, Perales HR (2010) Evolutionary response of landraces to climate change in centers of crop diversity. *Evol Appl* 2010 3:480–493
- Miller MP (2005) Alleles in space (AIS): computer software for the joint analysis of interindividual spatial and genetic information. *J Hered* 96:722–724
- Miller A, Schaal B (2005) Domestication of a Mesoamerican cultivated fruit tree, *Spondias purpurea*. *Proc Natl Acad Sci U S A* 102:12801–12806
- Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nat Rev Genet* 12:111–122
- Newton AC, Allnut TR, Gillies ACM et al (1999) Molecular phylogeography, intraspecific variation and the conservation of tree species. *Trends Ecol Evol* 14:140–145
- Odong TL, van Heerwaarden J, Jansen J et al (2011) Statistical techniques for defining reference sets of accessions and microsatellite markers. *Crop Sc* 51 doi:10.2135/cropsci2011.02.0095.
- Ouborg NJ, Pertoldi C, Loeschcke V et al (2010) Conservation genetics in transition to conservation genomics. *Trends Genet* 26:177–187
- Pascual L, Perfectti F, Gutierrez M, Vargas AM (1993) Characterizing isozymes of Spanish cherimoya cultivars. *HortScience* 28:845–847
- Palmberg-Lerche C (2008) Thoughts on the conservation of forest biological diversity and forest tree and shrub genetic resources. *J Trop For Sci* 20:300–312
- Perfectti F, Pascual L (1998) Characterization of cherimoya germplasm by isozyme markers. *Fruit Varieties J* 52:53–62
- Perfectti F, Pascual L (2005) Genetic diversity in a worldwide collection of cherimoya cultivars. *Genet Resour Crop Ev* 52:959–966
- Perry L (2012) Ethnobotany. In: Russo VM (ed) Peppers, production and uses, CABI, pp 1–13
- Petit RJ, El Mousadik A, Pons O (1998) Identifying populations for conservation on the basis of genetic markers. *Conserv Biol* 12:844–855
- Petit RJ, Aguinagalde I, de Beaulieu JL, Bittkau C (2003) Glacial refugia: hotspots but not melting pots of genetic diversity. *Science* 300:1563–1565

- Petit RJ, Hu FS, Dicks CW (2008) Forests of the past: a window to future changes. *Science* 320:1450–1452
- Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecol Model* 190:231–259
- Pinhasi R, Fort J, Ammerman AJ (2005) Tracing the origin and spread of agriculture in Europe. *PLoS Biol* 3:2220–2228
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Popenoe W (1921) The native home of the cherimoya. *J Hered* 12:331–336
- Popenoe H, King SR, León J et al (1989) Cherimoya. In: *Lost crops of the Incas: little-known plants of the Andes with promise for worldwide cultivation*. National Academy Press, Washington, DC, pp 228–239
- Ræbild A, Larsen AS, Jensen JS et al (2011) Advances in domestication of indigenous fruit trees in the West African Sahel. *New Forest* 41:297–315
- Ramachandran S, Deshpande O, Roseman CC et al (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 102:15942–15947
- Reed DH, Frankham R (2003) Correlation between fitness and genetic diversity. *Conserv Biol* 17:230–237
- Russell J, Dawson IK, Flavell AJ et al (2011) Analysis of > 1000 single nucleotide polymorphisms in geographically matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes. *New Phytol* 191:564–578
- Russell J, van Zonneveld M, Dawson IK et al (2013) Genetic diversity and ecological niche modelling of wild barley: refugia, large-scale post-Igm range expansion and limited mid-future climate threats. *PLoS ONE* under review
- Samuel AF, Drucker AG, Andersen SB et al (2013) Development of a cost-effective diversity-maximizing decision-support tool for in situ crop genetic resources conservation: The case of Cacao. *Ecol Econ* under review
- Scheldeman X, Van Damme P, Urea Alvarez JV, Romero Motoche JP (2003) Horticultural potential of Andean fruit crops exploring their centre of origin. *Acta Hort* 598:97–102
- Scheldeman X, van Zonneveld M (2010) Training manual on spatial analysis of plant diversity and distribution. *Biodiversity International*, Rome, Italy
- Schueler S, Kapeller S, Konrad H et al (2012) Adaptive genetic diversity of forest trees: promise for future forests and a threatened resource—a case study on Norway spruce in Austria. *Biodivers Conserv*. doi:10.1007/s10531-012-0313-3
- Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 227:1063–1066
- Thomas E, van Zonneveld M, Loo J et al (2012) Present spatial diversity patterns of *Theobroma cacao* L. in the Neotropics reflect genetic differentiation in Pleistocene refugia followed by human-influenced dispersal. *PLoS One* 7:e47676
- Tuberosa R, Graner A, Varshney RK (2011) Genomics of plant genetic resources: an introduction. *Plant Genet Resour* 9:151–154
- Van Damme P, Scheldeman X (1999) Promoting cultivation of cherimoya in Latin America. *Unasylva* 198:43–47
- Van Damme V, Gómez-Paniagua H, De Vicente MC (2010) The GCP molecular marker toolkit, an instrument for use in breeding food security crops. *Mol Breeding* 28:597–610
- van de Wouw M, Kik C, van Hintum T et al (2010a) Genetic erosion in crops: concept, research results and challenges. *Plant Genet Resour* 8:1–15
- van de Wouw M, van Hintum T, Kik C et al (2010b) Genetic diversity trends in twentieth century crop cultivars: a meta analysis. *Theor Appl Genet* 120:1241–1252
- van Etten J, Hijmans RJ (2010) A geospatial modelling approach integrating archaeobotany and genetics to trace the origin and dispersal of domesticated plants. *PLoS One* 5:e12060
- van Heerwaarden J, Hellin J, Visser RF, Eeuwijk FA van (2009) Estimating maize genetic erosion in modernized smallholder agriculture. *Theor Appl Genet* 119:875–888

- Vanhove W, Van Damme P (2009) Marketing of cherimoya in the Andes for the benefit of the rural poor and as a tool for agrobiodiversity conservation. *Acta Hort* 806:497–504
- Vanhove W, Van Damme P (2013) On-farm conservation of cherimoya (*Annona cherimola* Mill.) germplasm diversity. A value chain perspective. *Trop Conserv Sci* 6: 158–310
- van Zonneveld M, Thomas E, Galluzzi G, Scheldeman X (2011) Chapter 15/16: Mapping the ecogeographic distribution of biodiversity and GIS tools for plant germplasm collectors. In: Guarino L, Ramanatha RV, Goldberg E (eds) *Collecting Plant Genetic Diversity: Technical Guidelines—2011 Update*, Bioversity International, Rome, Italy (http://cropgenebank.sgrp.cgiar.org/index.php?option=com_content&view=article&id=662)
- van Zonneveld M, Scheldeman X, Escribano P et al (2012) Mapping genetic diversity of Cherimoya (*Annona cherimola* Mill.): application of spatial analysis for conservation and use of plant genetic resources. *PLoS One* 7:e29845
- Viruel MA, Hormaza JI (2004) Development, characterization and variability analysis of microsatellites in lychee (; Sonn., Sapindaceae). *Theor Appl Genet* 108:896–902
- Vinceti B, Loo J, Gaisberger H, et al (2013) Conservation priorities for *Prunus africana* defined with the aid of spatial analysis of genetic data and climatic variables. *PLoS ONE* 8: e59987
- Vranckx G, Jacquemyn H, Muys B, Honnay O (2011) Meta-analysis of susceptibility of woody plants to loss of genetic diversity through habitat fragmentation. *Conserv Biol* 26:228–237
- Waltari E, Hijmans RJ, Peterson AT et al (2007) Locating Pleistocene refugia: comparing phylogeographic and ecological niche model predictions. *PLoS One* 2:e563
- Widmer A, Lexer C (2001) Glacial refugia: sanctuaries for allelic richness, but not for gene diversity. *Trends Ecol Evol* 16:267–269
- Willis F, Moat J, Paton A (2003) Defining a role for herbarium data in Red List assessments: a case study of *Plectranthus* from eastern and southern tropical Africa. *Biodivers Conserv* 12:1537–1552
- Wolters B (1999) Zur Verbreitungsgeschichte und Ethnobotanik indianischer Kulturspflanzen, insbesondere des Kakaobaumes. *Angew Bot* 73:128–137
- Worthington M, Soleri D, Aragón-Cuevas F, Gepts P (2012) Genetic composition and spatial distribution of farmer-managed *Phaseolus* bean planting: an example from a village in Oaxaca, Mexico. *Crop Sc* 52:1721–1735

Chapter 5

Historical and Prospective Applications of ‘Quantitative Genomics’ in Utilising Germplasm Resources

Adrian Hathorn and Scott C. Chapman

Contents

5.1	Introduction	94
5.2	The Pedigree Era	94
5.2.1	The Infinitesimal Model	94
5.2.2	The Concept of Breeding Values	95
5.2.3	Selection Indices	95
5.2.4	Best Linear Unbiased Prediction (BLUP)	96
5.3	The Molecular Era	96
5.3.1	QTL Mapping	97
5.3.2	The Candidate Gene Approach	98
5.3.3	Gene Introgression and QTL Pyramiding	98
5.4	The Genomic Era	100
5.4.1	Genome-Wide Selection	100
5.4.2	Stepwise Regression, BLUP and the Bayesian Alphabet	101
5.4.3	How Many Markers Do We Need?	102
5.4.4	The Use of Low Density SNP Chips	103
5.4.5	Training Population Size and Design	103
5.4.6	Marker Assisted Recurrent Selection (MARS)	104
5.4.7	Maintaining Genetic Diversity	104
5.5	GWS or MAS/MARS?	106
	References	107

Abstract The last 30 years have seen major changes in the field of plant breeding. In this relatively short time frame we have witnessed the transition from a solely pedigree-based approach to genetic improvement, to one based almost entirely on genome-wide sequence information. We have also witnessed the evolution of dominant genetic theory, including the adoption of new statistical techniques necessary to accommodate the plethora of genomic information now available. In this chapter we review the past, present and future of plant breeding in terms of the three distinct “eras”: “the pedigree era”, “the molecular era” and the “genomics era”.

A. Hathorn (✉) · S. C. Chapman
CSIRO Plant Industry, Queensland Bioscience Precinct, 306 Carmody Rd.,
St. Lucia, QLD 4067, Australia
e-mail: adrian.hathorn@csiro.au

Keywords MAS · MARS · GWS · QTL mapping · Plant breeding · Genomics

5.1 Introduction

High-throughput genotyping technologies, in particular the single nucleotide polymorphism (SNP) chip, have prompted a revolution in the field of genetics and breeding. With the potential of genotyping literally hundreds of thousands of molecular markers at an affordable price, the once distant prospect of establishing an individual's genetic value without need of its pedigree has now become a reality. In the following chapter we consider the three distinct eras of quantitative genetics that led to this genotyping revolution, with final emphasis on genome-wide selection (GWS). We begin with “the pedigree era”, to describe analysis prior to DNA markers and revisit the fundamentals of quantitative genetic theory. In “the molecular era” we review the birth of molecular markers and the advent of marker assisted selection (MAS). Finally, in “the genomic era”, we describe the development of GWS and its current and future implications for plant breeding and utilization of genetic resources.

5.2 The Pedigree Era

5.2.1 *The Infinitesimal Model*

Prior to the era of molecular markers and genomics, genetic variation in quantitative traits was explained by modeling an individual's phenotype as the sum of an infinite number of infinitesimally small genetic effects plus an interaction between genotype and environmental values:

$$y_{ij} = \mu + g_i + e_{ij} \quad (5.1)$$

where; y_{ij} is the phenotype of individual i observed in environment j , μ refers to the fixed environmental effects of individual i , g_i is the total genetic value of individual i , and e_{ij} is the sum of random environmental effects affecting individual i in environment j . This is more commonly known as the infinitesimal model. The total genetic value of an individual g_i can be further partitioned into additive (g_A), dominance (g_D) and epistatic (g_E) components, with g_D and g_E representing the non-additive component of the genetic variation. Elaborations on this model include interaction effects of genotype with environment and also aim to consider interactions of different trait phenotypes using selection indices, as discussed below.

5.2.2 The Concept of Breeding Values

Substantial genetic improvement in animal breeding has been achieved by selecting on estimated breeding values (EBVs). However it has not been a popular index in plant breeding due mainly to the fact that the EBV is based on a simplified definition of heritability tailored towards selection on individual animals and does not take into account the diversity of observational units and mating systems used in plant breeding (Holland et al. 2010). However, EBVs are central to applications of genomic selection and are introduced here for that purpose.

The EBV represents the sum of the additive effects of an individual’s genes (Falconer and Mackay 1996; Lynch and Walsh 1998) and is typically used to determine an animal’s genetic potential when used as a parent. In its simplest form an EBV is estimated using the individual’s phenotype and the population narrow sense heritability, calculated as $h^2 = \sigma_A^2/\sigma_P^2$. The difference between an individual’s phenotypic value and the mean value of its population is adjusted according to h^2 in the following way:

$$EBV_i = m_0 + h^2(y_i - m_0) \quad (5.2)$$

where y_i is the phenotypic value of individual i and m_0 is the mean phenotypic value of the population. In this case, adjusting the phenotype according to the population narrow-sense heritability is a way of recognising that only a fraction of an individual’s phenotype is heritable. Furthermore, as each parent contributes a sample half of its genes to its progeny, it can also only transmit one-half of its genetic value. Thus the expected breeding value of the offspring of parents 1 and 2 is equal to:

$$E(\text{Progeny}_{P_1 \times P_2}) = m_0 + \frac{1}{2}EBV_{P_1} + \frac{1}{2}EBV_{P_2} \quad (5.3)$$

5.2.3 Selection Indices

Selection indices are a way of combining information across pedigrees and across traits. Equation 5.3 shows how the expectation of a breeding value can be defined using parental EBV’s. Depending on the accuracy of those breeding values, it may be useful to include information from more distant relatives such as grandparents. A selection index allows us to use this information in the one prediction thereby increasing the accuracy of genetic evaluation (Falconer and Mackay 1996).

The simplest example of using a selection index is the calculation of EBV’s based on own performance (Eq. 5.2). This takes the form of $I = EBV_i = b_{AP}P_i$ where b_{AP} is the simple regression of breeding value (A) on phenotype (P), and in the absence of any interaction between genotype and environment (GxE) is equal to $cov(A, P)/\sigma_P^2 = \sigma_A^2/\sigma_P^2 = h^2$ (Falconer and Mackay 1996). A selection index including information from a number of relatives therefore corresponds to a multiple

regression of breeding value on all the sources of information and the linear index of any one individual becomes:

$$I = b_{AP:1}P_1 + b_{AP:2}P_2 + b_{AP:3}P_3 + \dots \quad (5.4)$$

5.2.4 Best Linear Unbiased Prediction (BLUP)

A limitation of the selection index approach is that the method does not adjust the data for fixed environment effects; this must be done separately before the analysis. Henderson (1976) devised an efficient method to simultaneously estimate genetic and environmental effects in a single analysis. Henderson's method, called best linear unbiased prediction (BLUP), uses a mixed model approach and rapidly became the most widely accepted method of genetic evaluation due to its many desirable statistical properties.

The mixed model approach to estimating breeding values consists of modeling each trait value as the sum of all fixed environmental effects and a residual component comprised of the sum of all random genetic effects (BV's) plus a random error. In matrix notation,

$$y = X\beta + Za + e \quad (5.5)$$

where y is a vector of trait values, β is a vector of fixed environmental effects with incidence matrix X , a is a vector of random genetic effects with incidence matrix Z and e is a vector of errors. In plant breeding, alternative formulations of Eq. 5.5 are commonly used to incorporate random genotype by environment interaction effects as variance-covariance structures (Piepho 1997).

The basic mixed model used in both animal and plant breeding incorporates information from all relatives with or without phenotypic records to estimate BV's. Henderson (1976) showed that the expectation of fixed environmental effects $\hat{\beta}$ and the expectation of random genetic effects \hat{a} , are solutions to the mixed model equations:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \quad (5.6)$$

where $\hat{\beta}$ and \hat{a} are referred to as the best linear unbiased estimate (BLUE) and best linear unbiased predictor (BLUP), respectively.

5.3 The Molecular Era

The discovery of DNA markers marked the dawn of a new era in agricultural breeding, with some expectation that the ability to select directly on genotypes (MAS—marker assisted selection) would lead to the redundancy of pedigree based selection methods.

While DNA-based markers have now been deployed extensively for the tracking and introgression of simple traits (Eathington et al. 2007), their use in the selection for complex traits such as crop yield has so far been largely constrained to pyramiding of quantitative trait loci (QTL) in single cross populations.

The first DNA-based genetic markers were restriction fragment length polymorphisms (RFLPs) (Botstein et al. 1980), first used for the improvement of qualitative traits in crops by Beckmann and Soller (1986). Other early generation markers were enhanced by the introduction of DNA amplification-based procedures, such as; random amplified polymorphic DNAs (RAPDs) (Williams et al. 1990), AFLPs (Vos et al. 1995), as well as another broad class of DNA markers categorized as simple sequence repeats (SSRs) (Akkaya et al. 1992). SSR's are typically the most widely used markers in major cereals as they are highly reliable, co-dominant in inheritance and highly polymorphic (Collard and Mackill 2008).

5.3.1 *QTL Mapping*

One of the first applications of molecular marker technology was the discovery and mapping of quantitative trait loci (QTL). There are two distinct approaches to finding and mapping QTL. The first involves testing for marker-trait associations in a segregating population using marker genotypes located across the entire genome. The second, commonly referred to as the candidate gene approach, intuitively proposes previously sequenced genes of known function as potentially containing molecular polymorphisms related to the trait of interest.

The discovery and mapping of QTL in plants via marker trait associations, typically starts with the development of a mapping population, say 100 to 500 segregating individuals derived from an F2 or backcross population. Individuals (or, for hybrid crops, test-cross progeny) are then phenotyped for each trait of interest and genotyped with evenly spaced markers across the genome (linkage mapping). A variation on this is association mapping, where the individuals to be mapped represent a diverse set of relevant germplasm, e.g. historical (founder) and current breeding lines and potential donor lines for useful traits (Lynch and Walsh 1998).

The standard process of establishing significant marker trait associations is to use ordinary least squares where markers are treated as fixed effects and selected for inclusion into a prediction model using a stepwise regression approach based on arbitrary significance thresholds (Lande and Thompson 1990). The effects of markers below this threshold are set to zero, whilst those above the threshold are included in the model. The stepwise approach is useful to the extent that it minimizes the complexity of the model and ensures that there remain sufficient degrees of freedom for the estimation of marker effects. Once significant marker-trait associations have been made and major QTL identified, in theory at least, these major QTL are then ready to be introgressed into elite germplasm, hopefully leading to the development of new and improved cultivars.

One variation of this approach involves the genotyping of only that part of the population exhibiting extreme phenotypes for the target trait, creating two distinct

pools of DNA (Michelmore et al. 1991). Association is then inferred by finding allelic frequency differences between the groups of plants with contrasting phenotypes (Lebowitz et al. 1987). This approach is also referred to as “bulked segregant analysis” and has been successful in genetic mapping in plants using RFLP and SSR markers (Xu and Crouch 2008). Despite numerous reports for single major genes and major QTL using this method, bulked DNA analysis reports have been challenged by problems relating to insufficient marker density, low power of QTL detection and high false positive rates for marker-trait associations (Xu and Crouch 2008). See Van Eeuwijk et al. (2010) for a summary of analysis methods to derive performance estimates from different populations and statistical models.

5.3.2 The Candidate Gene Approach

The candidate gene approach has three chronological steps (Pflieger et al. 2001). First, candidate genes are proposed based on molecular and physiological studies of a trait. Then, a molecular polymorphism is identified so that statistical correlations between candidate gene polymorphisms and phenotypic variation can be calculated in a set of genetically unrelated individuals. The final step is the validation step and involves conducting complementary experiments to confirm the actual involvement of the candidate gene in the trait variation.

Although the candidate gene approach has been successfully used to characterize disease resistance genes and has led to the isolation of many new putative function resistance genes (R-genes), it is generally regarded as an expensive alternative to QTL mapping, especially for complex growth related traits. A recognized problem with the approach is that there are often a large number of candidate genes affecting a trait, so many genes must be sequenced in many individuals. The cost of carrying out so many association studies in a large sample of individuals is both expensive and time consuming. Furthermore, there is always a chance that the true causative mutation(s) may lie in a gene that would not intuitively have been selected as a candidate gene (Pflieger et al. 2001).

5.3.3 Gene Introgression and QTL Pyramiding

One of the more successful applications of molecular marker technology is in the introgression of major genes via marker assisted backcrossing (MABC). Although backcrossing has been successfully used in plant breeding to integrate disease resistance into numerous crop species such as maize (Hooker 1977) and wheat (Sharma and Gill 1983), prior to the invention molecular markers it was often a slow and complicated process. Without markers, phenotypic selection had to be done at each stage of the process. Fortunately, the implementation of MABC circumvented much of this process by using marker information to track target alleles from the donor parent

(Lamkey and Lee 2006). Large crop breeding programs, such as those for soybean and maize, have been redesigned to accommodate a seven-fold increase in data and analysis demands to implement accelerated breeding based on markers (Eathington et al. 2007). The same authors describe the complexities of using MABC to transfer transgene segments into multiple adapted genetic backgrounds.

A major problem in traditional backcrossing techniques is linkage drag. The identification of plants possessing the target trait and a high level of resemblance to the recurrent parent was complicated due to the fact that unfavorable alleles closely flanking the target allele would often “hitch a ride” into the recurrent parent. The implementation of MABC significantly helped researchers manage linkage drag by using marker information to identify plants with a high proportion of desirable genome from the recurrent parent. There are many examples of the successful use of MABC in rice, in particular bacterial blight resistance (Chen et al. 2000) and submergence tolerance (Toojinda et al. 2005), but also in other cereals such as barley, maize and wheat (see Table 1 in Collard and Mackill (2008)). This was further supplemented by simulation studies which found methods to optimize the recovery of recurrent parent alleles in just a few generations of backcrossing (Hillel et al. 1990; Hospital and Charcosset 1997; Visscher et al. 1996). For most crops, over 90 % of the recurrent parental genotype can now be recovered within two generations (Xu and Crouch 2008).

MABC is usually conducted in conjunction with phenotypic selection for other adaptive traits such as yield. However it remains a difficult challenge to introgress and pyramid multiple genes into a single cultivar even if they can be identified. Most plant breeders are forced to juggle selection on multiple traits and this often involves the more difficult process of selecting for many QTL simultaneously. The complexity of this task increases exponentially with increasing numbers of QTL. For example, if the frequency of 10 favorable alleles between two inbred parental lines is 0.5, then assuming they are unlinked, the frequency of the ideotype in the progeny will be equal to 1 in every 1024 recombinants. Pyramiding genes from more than two parents is an even tougher challenge. If those same 10 QTL were evenly distributed amongst three parents, the frequency of the ideotype would be around 1 in 60,000 recombinants! Some notable examples of successful QTL pyramiding in cereals include bacterial blight resistance in rice (Huang et al. 1997) and yellow mosaic virus in barley (Okada et al. 2004).

A common method to increase the frequency of target genotypes is through ‘enrichment’ of target alleles in segregating generations (e.g. F_2), followed by inbreeding (Bonnett et al. 2005; Wang et al. 2007). If the frequency of the ideotype is rare in early generations, a compromise is to select on heterozygotes at some, or all of the loci in the F_2 generation. By selecting for both target homozygotes and heterozygotes, this filter removes non-target homozygotes from the population. Through further inbreeding, or the development of doubled-haploid (DH) populations, the frequency of target homozygotes in the population can be increased along with the frequency of the target ideotype. The benefits of F_2 enrichment have also been demonstrated through simulation (Wang et al. 2009).

5.4 The Genomic Era

The molecular era was characterized by extensive searches for individual QTL using early generation markers such as RFLPs, RAPDs, AFLPs and SSRs. We also witnessed the advent of MAS. Consequently, substantial effort was directed to issues of marker density, population size, selection fractions and the combining of QTL across different genetic backgrounds.

In a sense, the transition to the genomic era came about largely through technological necessity. Most of the early generation markers were developed using the Sanger sequencing method (Sanger et al. 1978), which was both expensive and labor intensive. With the discovery of single nucleotide polymorphism (SNP¹) markers, there was increasing interest in developing a high-throughput low-cost assay that could make use of the relative abundance of these markers in both animal and plant genomes. A technological breakthrough came in the form of high density oligonucleotide arrays and quickly led to the development of massively parallel sequencing platforms, otherwise known as next-generation sequencing (NGS) platforms. The versatility of these arrays also allowed for the development of novel marker systems like single feature polymorphisms (SFPs), diversity array technology (DArT) and restriction site-associated DNA (RAD) markers (Gupta et al. 2008).

Although the emergence of NGS technologies has significantly reduced the cost of marker scoring (Shendure and Ji 2008), the development of new markers still requires significant investment (Deschamps et al. 2012). This is especially the case for crop species such as maize (SanMiguel et al. 1996) and wheat (Li et al. 2004), where the efficiency of the SNP discovery process is often hampered by large numbers of repetitive sequences. However, a new concept called genotype-by-sequencing (GBS) is beginning to emerge whereby massively parallel sequencing platforms are used to simultaneously develop and score SNP markers within a segregating population (Elshire et al. 2011). Since GBS can also be performed through a reduced representation approach (Van Orsouw et al. 2007), polymorphism discovery in larger and more complex genomes (e.g. allotetraploid durum wheat) is now becoming a simpler and more cost effective process (Trebbi et al. 2011).

5.4.1 Genome-Wide Selection

In 2001, (Meuwissen et al. 2001) proposed a method called “genome wide selection” or GWS, a simplification of the two step model selection approach detailed in (Lande and Thompson 1990). Rather than selecting a subset of markers for inclusion into the prediction model based on arbitrary significance thresholds, GWS proposed to exploit linkage disequilibrium (LD) within the genome by using all marker information in a single step to estimate individual genomic breeding values (GEBVs).

¹ SNP markers are point mutations commonly occurring throughout plant and animal genomes, whereby alleles differ by only one base position.

The implementation of GWS requires that a population of individuals (in structured or unstructured populations) be initially phenotyped for the trait of interest and genotyped for a pre-defined set of markers. This is referred to as the “training population”. The purpose of the training population is to accurately calibrate the prediction model by correlating marker effects with phenotypic values. The markers can then be used to estimate the genetic value of successive generations of individuals without need of phenotyping.

The method is theoretically superior to MAS for several reasons. The traditional MAS approach of fitting only the largest QTL is subject to a degree of upward bias known as the Beavis effect, an unavoidable consequence of selecting *a posteriori* among many estimates (Beavis 1998; Xu 2003). Lande and Thompson (1990) proposed a method to avoid this bias by using one half of the data to select the loci with the largest effects, and the other half to re-estimate the effects, although this was deemed to be a suboptimal use of the information (Meuwissen et al. 2001).

Furthermore, by fitting all markers into the prediction model these analyses should capture all (or most) of the additive genetic variance. This is in contrast to traditional MAS, where the estimation of a subset of significant QTL results in only a portion of the genetic variance being captured (Goddard and Hayes 2007). One consequence of this is that there is a greater chance that the largest and most accurately estimated QTL will be fixed in the first cycle of selection, leaving insufficient residual variation to maintain genetic gain in the cycles thereafter (Moreau et al. 2004).

5.4.2 Stepwise Regression, BLUP and the Bayesian Alphabet

Although stepwise regression is the technique of choice for selecting and fitting QTL markers in MAS, the choice of statistical technique for fitting all markers simultaneously in GWS is the topic of continuous debate. In GWS analyses, the number of marker effects to estimate will almost always be greater than the number of records (Goddard and Hayes 2007), and estimating a large number of marker effects in a data set of limited size leads to the problem of there not being enough degrees of freedom to fit all of the effects simultaneously via ordinary least squares (OLS). This is sometimes referred to as the ‘large p, small n’ problem. If the significance thresholds in stepwise regression are sufficiently relaxed, thereby allowing for a greater spectrum of QTL effects (e.g. additive x additive interactions) to be included in the model, it has been shown that prediction accuracies as high as 0.61 can be achieved using this method (Habier et al. 2007). This is still not ideal however, since in order to exploit the full potential of GWS we must make use of all available marker information.

The only way to use all marker information is to treat the markers as random effects within a BLUP or Bayesian framework. Meuwissen et al. (2001) used simulation to compare the accuracy of GEBVs using ridge regression BLUP (RR-BLUP) and two Bayesian methods called BayesA and BayesB. Whilst RR-BLUP assumes that marker effects are normally distributed with constant variance (Whittaker et al.

2000), both BayesA and BayesB assume slightly different forms of an inverted Chi-square prior distribution of marker effects. Although RR-BLUP was highly accurate ($\gamma_{TBV;EBV} = 0.73$) BayesA and BayesB were clearly superior ($\gamma_{TBV;EBV} = 0.80$ and 0.85 respectively), especially when QTL vary in magnitude. Other variations on the “Bayesian alphabet” have also been shown to be highly effective (Habier et al. 2011). For an in-depth discussion of Bayesian methodology in a breeding context see (Gianola et al. 2009).

More recently, Hayes et al. (2009a) proposed using dense marker information to predict realized relationship coefficients between pairs of individuals using BLUP. The method, which we will refer to as genomic BLUP (G-BLUP), is therefore analogous in principle to traditional pedigree BLUP in that it attempts to calculate in each case the proportion of the genome that is identical by descent (IBD). It is superior to pedigree BLUP because it calculates this value directly, rather than relying on an expectation derived through lineage and selection, and is therefore able to account for Mendelian sampling during gamete formation.

5.4.3 *How Many Markers Do We Need?*

Marker density affects both the prediction of individual marker effects in all forms of GWS and the estimation of realized relationship coefficients in G-BLUP. This is because the accuracy of prediction for both methods relies in large part, on the ability of markers to serve as proxies for QTL from generation to generation. So what is the ideal marker density?

Firstly, the ability of markers to act as proxies for QTL, is a function of average LD within the genome. The greater the expanse of LD in the genome, the fewer markers that will be required to tag QTL located in any particular region. Furthermore, for each generation that GWS is practiced, the proportion of genetic variance explained by each marker decreases and the accuracy of GWS will tend to decline for each successive generation that it is practiced (Muir 2007).

The rates of LD decay are known to vary considerable between species, depending on a range of population characteristics, including those affected by selection history (Gaut and Long 2003). From a genetic perspective, species LD will depend on the population recombination rate, $4N_e r$ where N_e is equal to the effective population size, and r is the recombination rate per base pair. If this is known, the target marker density for GWS can be approximated by using the average r^2 between adjacent markers as a measure of their marker density relative to the decay of LD (Calus and Veerkamp 2007; Heffner et al. 2009).

The importance of marker density can also be demonstrated in terms of the relationship between the effective population size N_e and the number of independent chromosome segments q , defined as $q = 2N_e \times L$ where L is equal to the total genomic map length in Morgans (Hayes et al. 2009b). Obviously if N_e is large, the number of ‘independent’ chromosome segments is also large, and the extent of LD in the population will be limited, requiring a very large number of markers to capture all QTL effects.

5.4.4 The Use of Low Density SNP Chips

Despite rapid advancements in genotyping technology, the cost of genotyping (including sample collection and DNA extraction) remains a potential limitation in the implementation of GWS in smaller breeding programs (Ibañez-Escriche and Gonzalez-Recio 2011). This is especially the case when the number of selection candidates per generation is high, or the economic benefit per selection candidate is low compared to the cost of genotyping (Habier et al. 2009).

One solution is to use a reduced set of markers in the selection candidates to reduce overall cost while minimising loss of accuracy. One adaptation of this strategy involves the use of variable selection methods to identify a small set of markers that are predictive of trait phenotype or breeding value. Although variable selection methods have been shown to have good predictive ability (Cleveland et al. 2010; Iwata and Jannink 2010; Vazquez et al. 2010; Weigel et al. 2009), this approach is less attractive for multiple trait selection and across populations since it requires specific SNPs for each trait and population (Ibaez-Escriche and Gonzalez-Recio 2011).

A more flexible (but less accurate) approach was proposed by Habier et al. (2009) who suggested using evenly spaced low-density markers to obtain GEBVs in the selection candidates. This way, co-segregation of high and low density SNPs within families can be used to impute missing SNP genotypes in the selection candidates. The method is similar to the use of TAG SNPs to identify haplotype blocks segregating across human populations (Servin and Stephens 2007), the primary difference being that Habier et al. (2009) used pedigrees to estimate haplotype blocks within families rather than across populations.

5.4.5 Training Population Size and Design

Another avenue to reduce genotyping costs is to limit the size of the training population through selective genotyping. Selective genotyping has been previously proposed to improve the efficiency of QTL detection in a linkage mapping context (Sun et al. 2010) and more recently in a GWS context (Zhao et al. 2012). Genotyping only those candidates with high or low phenotypic values of the target trait (bidirectional selection) has been shown to lead to only a marginal decrease in the prediction accuracy of genomic breeding values (Zhao et al. 2012) and may therefore be a useful way to conduct GWS under a restricted budget.

Further concessions can also be made in crops when conducting GWS within large bi-parental populations, since bi-parental populations have extensive LD and allow for complete genome coverage with only a few hundred markers (Heffner et al. 2011). In bi-parental GS, the training population is made up of a subset of the progeny and the resulting prediction models are then used for predicting genetic value of the remaining progeny or for subsequent cycles of marker assisted recurrent selection (Bernardo and Yu 2007).

5.4.6 *Marker Assisted Recurrent Selection (MARS)*

Compared to crop breeders, animal breeders have been more accepting of GWS, perhaps due to a long history of selecting on breeding values as a surrogate for aggregate performance (Nakaya and Isobe 2012). In actual fact, the concept of selecting on an index is not at all foreign to many modern plant breeders. Many breeders practice a ‘simplified’ version of GWS referred to as “marker assisted recurrent selection” or MARS (Eathington et al. 2007; Hospital and Charcosset 1997; Koebner 2003). Like GWS, MARS focuses on individual performance rather than marker genotypes and selects on “marker scores” rather than GEBVs.

MARS was made possible by Lande and Thompson (1990) who first derived optimal selection indices for the improvement of quantitative traits, using both molecular and phenotypic information. Simply put, MARS refers to the improvement of typically an F_2 population by a single cycle of MAS (based on phenotypic and marker scores) followed by three cycles of selection based on marker scores only (Bernardo and Yu 2007). Thus there are two distinct steps involved in the application of MARS: model selection and model estimation. The model selection step involves identifying F_2 or F_2 derived progeny with a high proportion of favorable alleles at target marker loci. Markers are typically chosen on the basis of a statistical test for significance. In the model estimation step, each marker is given a weight based on its estimated effect and individual candidates are then ranked based on selection index (molecular score). The selection index represents a prediction of genetic value of each line, much the same as a GEBV.

The primary benefit of MARS, as opposed to traditional MAS, is that once the linear model for estimation of breeding values has been derived, it can be used to predict the breeding value of other marker genotyped individuals within the population or in future generations. MARS has been widely used in plant breeding programs in bi-parental crosses (Eathington et al. 2007) but is now being replaced by approaches that utilize data compiled across the entire breeding program. As LD in successive generations is slowly eroded due to recombination, the prediction value of each marker decreases. With tight linkage between marker and trait loci, the breakdown of LD can be minimized and the model can be used for several cycles of selection. Theoretically, if all markers were to be attributed the same effect, MARS would actually be equivalent to F_2 enrichment (Bernardo 2008). However since there is variation in marker effects, an individual with two QTL with large effects can be prioritized for selection in MARS ahead of an individual with four QTL with small effects.

5.4.7 *Maintaining Genetic Diversity*

The loss of genetic diversity in elite breeding programs due to factors such as selection, small population sizes and genetic drift remains an issue for concern for plant breeders. The importance of genetic diversity can be thought of in terms of its role

in generating additive genetic variance (σA) and thus genetic gain. It may be useful to think of the additive genetic variance as stored genetic potential, and selection as being the process through which this genetic potential is converted to genetic gain. This can be seen in the following equation:

$$R = ih^2\sigma_p$$

where R = response to selection, i = intensity of selection, h^2 = heritability, equal to σ_A/σ_p and σ_p equals the phenotypic standard deviation (Falconer and Mackay 1996).

Selection can alter the genetic variance in a population by either changing allele frequencies and/or generating linkage disequilibrium (Lynch and Walsh 1998). However depending on the type of selection being applied to the population, the genetic variance may either increase due to the generation of coupling disequilibrium (e.g. disruptive selection), or decrease due to the generation of repulsion disequilibrium (e.g. directional and stabilizing selection). This reduction in additive genetic variance, otherwise known as the Bulmer effect (Bulmer 1976), has been shown to adversely affect the response to both GWS and pedigree based BLUP selection (Van Grevenhof et al. 2012). The challenge for breeders is therefore to practice selection whilst preserving genetic variance or risk severely limiting their opportunities to force further adaptation in the long term.

One approach to this issue involves the introgression of new alleles for quantitative traits from unadapted germplasm. Plant breeders have been understandably hesitant of this approach primarily due to the risk of inadvertently breaking up favorable linkage blocks in elite germplasm. Many of these linkage blocks have been formed through the gradual accumulation of favorable genes linked in coupling over many generations. Reassembling favorable linkage blocks can be difficult, especially for small effect genes whose positive effects can be masked by the negative effects of other linked genes introduced from the unadapted parent (Jordan et al. 2011).

With the availability of high-density marker platforms to better understand the genetic architecture of both donor and recipient populations (Klein et al. 2008), it is becoming increasingly possible to utilize crossing strategies such as nested association mapping (Buckler et al. 2009; Yu et al. 2008) and modified backcrossing strategies (e.g. Jordan et al. 2011) to more quickly and efficiently deliver novel genetic segments into adapted cultivars. Indeed, high-density marker platforms are now being used extensively for assessments of genetic diversity in cereals (Chen et al. 2012; Pan et al. 2012), as well as in the discovery of allelic variants of known genes (Deschamps and Campbell 2010). Furthermore, with the ability to select directly on high-density markers, GWS allows for the option of preferentially weighting low frequency favorable markers in the early cycles of selection so as to avoid losing low frequency favorable QTL. Simulation has shown this to be a potentially useful way of maximizing response over the long term, whilst sacrificing little or no response in the short term (Jannink 2010).

5.5 GWS or MAS/MARS?

Since its inception in 2001, GWS has had a limited uptake in public plant breeding programs, although an internet search will show that large seed companies have been hiring many staff in these areas in the last 2-3 years. The slow uptake in public breeding is despite a growing body of evidence in both plants and animals suggesting that GWS is the most efficient way of making use of the availability of dense marker information. For example, Bernardo and Yu (2007) simulated the response due to GWS compared with MARS over three cycles of selection in maize and found the response due to GWS to be 18 to 43 % larger than the response due to MARS. Furthermore, through reduction in time and costs needed to prove the value of a bull, Schaeffer (2006) showed that GWS could provide a twofold increase in response to selection and save 92 % of the costs of the current progeny test based breeding programs. Finally, (Wong and Bernardo 2007) showed through simulation that the 19 year selection cycle in oil palm could be reduced to 6 years, and that GWS would outperform MARS and phenotypic selection on a gain per unit cost and time basis, even with very small population sizes ($N = 50$). It may be that some of the reticence to implement GWS is due to lack of demonstration of the realised results. Most of the published studies are related to prediction of target populations from training populations, rather than the realisation of recombination over several cycles of selection, and it will be some time before supporting evidence is accumulated in this area.

Further motivation for the adoption of GWS should lie in the fact that MAS has not performed up to the expectations set on it over two decades ago. Since the first application of molecular markers to crop improvement in 1986, an interesting dichotomy has developed between the number of publications purporting to have found significant marker-trait associations and the number of publications describing the successful development of finished breeding products. By the year 2000, the number of publications containing the term “quantitative trait loci” outnumbered those that contained the term “marker assisted selection” on Google Scholar by a factor of 3 (Xu and Crouch 2008), and the gap appears to have widened since. Thus although the discovery of marker trait associations in crops has been successful (Price 2006), the application of this knowledge into developing new plant varieties has not, at least in the public sector. Potential reasons for this disparity are presented by Collard and Mackill (2008).

MAS and its various adaptations should be viewed collectively as interim methodologies, developed to make the best use of limited available information at a time when genotyping technologies were very much in their infancy. Much has changed since this time. Over the last 20 years, the number of base pairs sequenced per dollar has increased exponentially, and by extension, so has the amount of genomic information available for analysis. With the cost of DNA sequencing now dropping by half every 5 months (Stein 2010), DNA sequencing throughput is outpacing advancements in both computer speed and storage capacity. A major limitation for smaller breeding programs is that they need efficient and well-designed information systems

together with skilled staff to exploit these opportunities. Given these capabilities are available, it is therefore opportune for plant breeding programs to transit from traditional and MAS breeding to GWS and strive to exploit this influx of genomic information.

Acknowledgment This publication has been partially supported by external funding from the Generation Challenge Program and the Bill and Melinda Gates Foundation as part of the Integrated Breeding Platform project which supports training and implementation of marker technologies and open-source tools in public breeding programs (<https://www.integratedbreeding.net/>).

References

- Akkaya MS, Bhagwat AA, Cregan PB (1992) Length polymorphisms of simple sequence repeat DNA in soybean. *Genetics* 132:1131–1139
- Beavis WD (1998) The power and deceit of QTL experiments: lessons from comparative QTL studies. 49th annual corn and sorghum industry research conference. ASTA, Washington, pp 145–162
- Beckmann JS, Soller M (1986) Restriction fragment length polymorphisms in plant genetic improvement. *Oxf Surv Plant Mol Cell Biol* 3:196–250
- Bernardo R (2008) Molecular markers and selection for complex traits in plants: learning from the Last 20 Years. *Crop Sci* 48:1649–1664
- Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090
- Bonnett DG, Rebetzke GJ, Spielmeyer W (2005) Strategies for efficient implementation of molecular markers in wheat breeding. *Mol Breed* 15:75–85
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* 32:314–331
- Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
- Bulmer MG (1976) The effect of selection on genetic variability: a simulation study. *Genet Res* 28:101–117
- Calus M, Veerkamp RF (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *J Anim Breed Genet* 124:362–368
- Chen S, Lin XH, Xu CG, Zhang Q (2000) Improvement of bacterial blight resistance of 'Minghui 63', an elite restorer line of hybrid rice, by molecular marker-assisted selection. *Crop Sci* 40:239–244
- Chen X, Min D, Yasir TA, Hu Y-G (2012) Genetic diversity, population structure and linkage disequilibrium in elite chinese winter wheat investigated with SSR markers. *PLoS ONE* 7:e44510
- Cleveland M, Forni S, Deeb N, Maltecca C (2010) Genomic breeding value prediction using three bayesian methods and application to reduced density marker panels. *BMC Proc* 4:S6
- Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos T T Soc A* 363:557–572
- Deschamps S, Campbell MA (2010) Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Mol Breed* 25:553–570
- Deschamps S, Llaca V, May GD (2012) Genotyping-by-sequencing in plants. *Biology* 1:460–483
- Eathington SR, Crosbie TM, Edwards MD et al (2007) Molecular markers in a commercial breeding program. *Crop Sci* 47:S154–S163

- Elshire RJ, Glaubitz JC, Sun Qi et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for High diversity species. *PLoS ONE* 6:e19379
- Falconer DS, Mackay TFC (1996) Introduction to Quantitative Genetics. Benjamin Cummings
- Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. *Plant Cell* 15:1502–1506
- Gianola D, De Los Campos G, Hill WG et al (2009) Additive genetic variability and the Bayesian alphabet. *Genetics* 183:347–363
- Goddard ME, Hayes BJ (2007) Genomic selection. *J Anim Breed Genet* 124:323–330
- Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. *Heredity* 101:5–18
- Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177:2389–2397
- Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using low-density marker panels. *Genetics* 182:343–353
- Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics* 12:186
- Hayes BJ, Visscher PM, Goddard ME (2009a) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* 91:47
- Hayes BJ, Daetwyler HD, Bowman P et al (2009b) Accuracy of genomic selection: comparing theory and results. In: Proceedings of the 18th conference: association for the advancement of animal breeding and genetics, Barossa Valley, South Australia, pp 34–37
- Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
- Heffner EL, Jannink J-L, Iwata H et al (2011) Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Sci* 51:2597–2606
- Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32:69–83
- Hillel J, Schaap T, Haberfeld A et al (1990) DNA fingerprints applied to gene introgression in breeding programs. *Genetics* 124:783–789
- Holland JB, Nyquist WE, Cervantes-Martínez CT (2010) Estimating and interpreting heritability for plant breeding: an update. In: Janick J (ed) *Plant breeding reviews*, Wiley, Oxford, pp 9–112
- Hooker AL (1977) A plant pathologist's view of germplasm evaluation and utilization. *Crop Sci* 17:689–694
- Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. *Genetics* 147:1469–1485
- Huang N, Angeles ER, Domingo J et al (1997) Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. *Theor Appl Genet* 95:313–320
- Ibañez-Escriche N, Gonzalez-Recio O (2011) Review promises, pitfalls and challenges of genomic selection in breeding programs. *Span J Agric Res* 9:404–413
- Iwata H, Jannink J-L (2010) Marker genotype imputation in a low-marker-density panel with a high-marker-density reference panel: accuracy evaluation in barley breeding lines. *Crop Sci* 50:1269–1278
- Jannink J-L (2010) Dynamics of long-term genomic selection. *Genet Sel Evol* 42:35
- Jordan DR, Mace ES, Cruickshank AW et al (2011) Exploring and exploiting genetic variation from unadapted sorghum germplasm in a breeding program. *Crop Sci* 51:1444–1457
- Klein RR, Mullet JE, Jordan DR et al (2008) The effect of tropical sorghum conversion and inbred development on genome diversity as revealed by high-resolution genotyping. *Crop Sci* 48:S12–26
- Koebner R (2003) MAS in cereals: green for maize, amber for rice, still red for wheat and barley. In: *Marker assisted selection: A fast track to increase genetic gain in plant and animal breeding?* Turin, Italy. 17–18 Oct 2003
- Lamkey KR, Lee M (2006) *Plant breeding: the Arnel R Hallauer international symposium*. Wiley-Blackwell
- Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756

- Lebowitz RJ, Soller M, Beckmann JS (1987) Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theor Appl Genet* 73:556–562
- Li W, Zhang P, Fellers JP, Friebe B, Gill BS (2004) Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J* 40:500–511
- Lynch M, Walsh B (1998) Genetics and analysis of quantitative Traits. Sinauer Associates
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *PNAS* 88:9828–9832
- Moreau L, Charcosset A, Gallais A (2004) Experimental evaluation of several cycles of marker-assisted selection in maize. *Euphytica* 137:111–118
- Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet* 124:342–355
- Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? *Ann Bot* 110:1303–1316
- Okada Y, Kanatani R, Arai S, Ito K (2004) Interaction between barley yellow mosaic disease-resistance genes *rym1* and *rym5*, in the response to BaYMV strains. *Breed Sci* 54:319–325
- Pan Q, Ali F, Yang X et al (2012) Exploring the genetic characteristics of two recombinant inbred line populations via high-density SNP markers in maize. *PLoS ONE* 7:e52777
- Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review. *Mol Breed* 7:275–291
- Piepho HP (1997) Analyzing genotype-environment data by mixed models with multiplicative terms. *Biometrics* 761–766
- Price AH (2006) Believe it or not, QTLs are accurate!. *Trends Plant Sci* 11:213–216
- Sanger F, Coulson AR, Friedmann T et al (1978) The nucleotide sequence of bacteriophage ϕ X174. *J Mol Biol* 125:225–246
- SanMiguel P, Tikhonov A, Jin YK et al (1996) Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765–768
- Schaeffer LR (2006) Strategy for applying genome-wide selection in dairy cattle. *J Anim Breed Genet* 123:218–223
- Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* 3:e114
- Sharma HC, Gill BS (1983) Current status of wide hybridization in wheat. *Euphytica* 32:17–31
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135–1145
- Stein L (2010) The case for cloud computing in genome informatics. *Genome Biol* 11:207
- Sun Y, Wang J, Crouch JH, Xu Y (2010) Efficiency of selective genotyping for genetic analysis of complex traits and potential applications in crop improvement. *Mol Breed* 26:493–511
- Toojinda T, Tragoonrung S, Vanavichit A et al (2005) Molecular breeding for rainfed lowland rice in the mekong region plant production. *Science* 8:330–333
- Trebbi D, Maccaferri M, Heer P de et al (2011) High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf). *Theor Appl Genet* 123:555–569
- Van Eeuwijk FA, Bink MC, Chenu K, Chapman SC (2010) Detection and use of QTL for complex traits in multiple environments. *Curr Opin Plant Biol* 13:193–205
- Van Grevenhof EM, Van Arendonk JA, Bijma P (2012) Response to genomic selection: the bulmer effect and the potential of genomic selection when the number of phenotypic records is limiting. *Genet Sel Evol* 44:26
- Van Orsouw NJ, Hogers RCJ, Janssen A et al (2007) Complexity reduction of polymorphic sequences (CRoPSTM): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* 2:e1172
- Vazquez AI, Rosa GJM, Weigel KA et al (2010) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. *J Dairy Sci* 93:5942–5949

- Visscher PM, Haley CS, Thompson R (1996) Marker-assisted introgression in backcross breeding programs. *Genetics* 144:1923–1932
- Vos P, Hogers R, Bleeker M et al (1995) AFLP: a new technique for DNA fingerprinting. *Nucl Acids Res* 23:4407–4414
- Wang J, Chapman SC, Bonnett DG et al (2007) Application of population genetic theory and simulation models to efficiently pyramid multiple genes via marker-assisted selection. *Crop Sci* 47:582–588
- Wang J, Chapman S, Bonnett D, Rebetzke G (2009) Simultaneous selection of major and minor genes: use of QTL to increase selection efficiency of coleoptile length of wheat (*Triticum aestivum* L). *Theor Appl Genet* 119:65–74
- Weigel KA, De Los Campos G, González-Recio O et al (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci* 92:5248–5257
- Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. *Genet Res* 75:249–252
- Williams JGK, Kubelik AR, Livak KJ et al (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl Acids Res* 18:6531–6535
- Wong C, Bernardo R (2007) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theor Appl Genet* 116:815–824
- Xu S (2003) Theoretical basis of the beavis effect. *Genetics* 165:2259–2268
- Xu Y, Crouch JH (2008) Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci* 48:391–407
- Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539–551
- Zhao Y, Gowda M, Longin F et al (2012) Impact of selective genotyping in the training population on accuracy and bias of genomic selection. *Theor Appl Genet* 125:707–713

Part II
Platforms and Approaches to Investigate
Plant Genetic Resources

Chapter 6

High-throughput SNP Profiling of Genetic Resources in Crop Plants Using Genotyping Arrays

Martin W. Ganal, Ralf Wieseke, Hartmut Luerssen, Gregor Durstewitz, Eva-Maria Graner, Joerg Plieske and Andreas Polley

Contents

6.1	Introduction	114
6.2	Identification of SNPs	115
6.2.1	Transcriptome Sequencing	115
6.2.2	Reduced Complexity Sequencing	116
6.2.3	Whole Genome Sequencing	117
6.3	Selection of SNPs for a Genotyping Array	117
6.4	SNP Calling Based on Array Data	119
6.5	Analysis of SNP Genotyping Data from a Large Array	122
6.6	Large SNP arrays in crop plants and examples for their use	122
6.6.1	Availability of Large Genotyping Arrays for Crop Plants	123
6.6.2	Examples for the Use of Large Genotyping Arrays for the Characterization of Plant Germplasm and Varieties	123
6.7	Summary and Future Trends	126
	References	127

Abstract Using high-throughput DNA sequencing technologies, it is now possible to quickly and reliably identify many thousands to millions of SNPs in a species. They can subsequently serve as markers for the development of large genotyping arrays. Large numbers of individuals derived from gene banks, landraces, breeding material and varieties can be genotyped with such arrays at an extremely high marker density in a fast, efficient and highly reproducible way. Based on our experience, we provide in this chapter an overview on various aspects that have to be considered within the process of developing such genotyping arrays, including the SNP discovery and/or collection, possible selection criteria for SNPs to be put on the array, SNP scoring and allele calling as well as data assembly for the analysis of millions of genotypes. To make the best use of these genotyping data, it will be very important to establish databases containing marker data from many genotyping experiments in order to simplify downstream data processing for scientific as for breeding purposes.

M. W. Ganal (✉) · R. Wieseke · H. Luerssen · G. Durstewitz · E.-M. Graner · J. Plieske · A. Polley
TraitGenetics GmbH, Am Schwabeplan 1b, 06466 Gatersleben, Germany
e-mail: ganal@traitgenetics.de

Keywords Single nucleotide polymorphism · Molecular marker · DNA sequencing · Maize · Tomato

6.1 Introduction

Over the last 20–25 years, genetic analyses using molecular markers have involved an ever-increasing number of molecular markers of different types. Starting with RFLP (restriction fragment length polymorphisms) markers and continuing through various types of PCR-based markers (e.g. AFLPs or other types of amplified fragment length polymorphisms), this has ultimately led to the analysis of single nucleotide polymorphisms (SNPs). SNPs are single base differences between individuals within an orthologous DNA sequence. Since SNPs are the smallest unit of any polymorphism type, they are most abundant in a given genome and the limitation on the availability of such polymorphisms is essentially based on the level of polymorphism in a species. In most organisms including crop plants, SNPs are now the cornerstone of genetic marker analysis.

Initially, molecular markers have been used to determine the genetic relationship between individuals and to generate low to medium density molecular marker data sets and genetic maps. These could then be used to localize monogenic (inherited by a single gene) or polygenic traits (influenced by a number of genes and also termed QTLs or quantitative trait loci). With larger numbers of scorable markers, it has been possible to map such loci with ever-increasing accuracy in segregating populations, thus opening the door for more complex applications.

Such research topics are the precise characterization of genetic material or populations with many thousands of markers so that detailed information about population structure, linkage disequilibrium and genetic diversity on a whole-genome level or for specific regions of the genome can be accumulated. This knowledge is necessary for so-called genome-wide association studies (GWAS). GWAS is an approach that allows the identification of chromosomal regions which have a significant effect on quantitative traits with high accuracy. It requires the availability and use of very large numbers of molecular markers (almost exclusively SNP markers) for scanning a genome. GWAS has originally mainly been used in humans for the identification of genes and chromosomal regions that were associated with complex diseases, but meanwhile it is also used in many other organisms for the identification of chromosomal regions associated with complex and quantitatively inherited traits (Rafalski 2010). Finally in animal and plant breeding, large scale genotyping with very large numbers of markers gains enormous interest within a procedure called genomic selection (GS). This procedure is applied to populations of unrelated individuals from a breeding pool in order to predict phenotypes exclusively based on genotype data (Meuwissen et al. 2001). Many markers are being used to estimate their individual effects, and their cumulative effects are used as predictors for the actual phenotype. GS is mainly based on the hypothesis that many small effects contribute to the final expression of the phenotype. Since linkage disequilibrium is relatively small in many

organisms, GS also requires genotyping of very large numbers of markers (Hamblin et al. 2011).

Genotyping of very large numbers (many thousands) of markers has been enabled through the development of technologies that permit the simultaneous analysis of many thousands of SNP markers on genotyping arrays. Currently, two main technologies are being used for that purpose. One technology (Affymetrix) is based on the use of solid phase bound oligonucleotide probes on an array and the subsequent hybridization of genomic DNA onto such arrays (McGall and Christians 2002). Another frequently used technology (Illumina Infinium) is based on the use of single base primer extension to determine the specific allelic state for a given single copy locus. Together with specific anchor and identifier sequences attached to the primer, large numbers of loci can be analyzed simultaneously on arrays or chips that contain the respective complementary sequences at specific positions (Gunderson et al. 2006; Steemers et al. 2006). For example in human SNP analyses, such arrays now permit the parallel genotyping of several millions of SNPs.

In plants, large SNP genotyping arrays are now being established for a number of species. In this chapter, we will review our the experiences during the design and analysis of a number of large genotyping arrays for species such as tomato, oilseed rape, barley, wheat and especially maize where we have been involved in the design and analysis of the largest published plant SNP genotyping array containing nearly 50,000 SNP markers (Ganal et al. 2011).

6.2 Identification of SNPs

A major prerequisite for the development of a large genotyping array is the availability of very large numbers of SNPs. Until a few years ago, it was quite laborious to identify many thousands of SNPs through methods such as amplicon sequencing (Ganal et al. 2009). This situation has changed now with the arrival of the novel or Next Generation Sequencing (NGS) technologies (Metzker 2010) that permit the efficient generation of hundreds of millions to billions of bases of sequencing information. With a single run, it is now possible to obtain sequences from the entire transcriptome or the full genome of plant species (Varshney et al. 2009), and the major challenge has moved from the sequencing process to the bioinformatic identification of SNPs in these vast sequence data sets. In crop plants, for the large scale of identification of SNPs, three different approaches are mainly used.

6.2.1 Transcriptome Sequencing

Transcriptome sequencing means the sequencing of the reverse transcribed mRNA fraction. This approach has been used in comparative transcriptome sequencing of different lines or varieties. One advantage of transcriptome sequencing is that the

number of transcribed genes in a plant species is around 20,000 to 40,000 independent of the genome size. Through the use of normalization procedures and the inclusion of different tissues, it is possible to obtain reliable sequencing information on the transcribed part of most genes (Iorizzo et al. 2011). Furthermore, since many genes are present only in single copy in a genome, SNPs identified in this genome fraction are more likely to be useful for marker development compared to other sequences obtained from duplicated or repetitive sequences. Transcriptome sequencing has been used in many different crop species to identify large numbers (many thousands) of SNPs through the comparison of contigs obtained for orthologous genes in different lines or varieties (Barbazuk et al. 2007; Novaes et al. 2008; Imelfort et al. 2009; Trick et al. 2009; Han et al. 2011; Blanca et al. 2012). A disadvantage of SNP identification in transcribed sequences may be that the number of SNPs in this fraction can be relatively low due to selection constraints, especially in organisms with a generally low level of polymorphism, and different contigs could be derived from the same gene. SNPs near exon/intron boundaries might not be useful for marker development due to the lack of the intron sequences and the distribution of SNPs in the genome might be uneven since low copy sequences in untranscribed regions of the genome cannot be analyzed. In addition, important regulatory switches (i.e. those that affect transcription) are located outside of mRNA transcripts (promoters, transcriptional enhancers).

6.2.2 *Reduced Complexity Sequencing*

The described potential problems of transcriptome sequencing have been circumvented in part by reducing the complexity of the genome before sequencing, especially in plant species with a very large genome. Complexity reduction of genomic DNA can be achieved in different ways. One way is the selection of undermethylated DNA through the digestion with methylation-sensitive restriction enzymes or combinations of such enzymes with other restriction enzymes (Gore et al. 2009a, b; Deschamps et al. 2010; Hyten et al. 2010). Other approaches include a modification of the AFLP procedure (CRoPS technology) or the RAD technology (van Orsouw et al. 2007; Barchi et al. 2011; Davey et al. 2011; Trebbi et al. 2011). Using these technologies before the actual sequencing, it is possible to selectively enrich for SNPs in predominantly single or low copy sequences including also the non-coding fraction of the genome. Due to the fact that it is difficult to clearly discriminate orthologous from paralogous sequences, these technologies require more sophisticated bioinformatic procedures to reliably identify SNPs, and the number of identified SNPs is usually only in the range of a few thousands to ten thousands.

Another more recent technology that reduces the complexity of the sequenced fraction of the genome uses a process called sequence capture. In this process, DNA sequences derived from single copy genes are used as baits placed on a microarray or in liquid phase to specifically capture these sequences. Since up to 50 Mbp of

sequence can be captured, it is in principle possible to capture all coding sequences and use only this fraction for SNP detection. First data on the complex genomes of maize and wheat suggest that this approach is feasible, a significant enrichment of gene sequences can be achieved, and SNPs can be identified (Fu et al. 2010; Saintenac et al. 2011; Winfield et al. 2012).

6.2.3 Whole Genome Sequencing

The most comprehensive procedure for the identification of SNPs in a species is the comparative sequencing of the entire genome and its comparison to a high quality reference sequence as they are available for the model species *Arabidopsis* and rice, both of which are relatively small in size (*Arabidopsis* Genome Initiative 2000; International Rice Genome Sequencing Project 2005). This approach permits the detection of nearly all sequence polymorphisms in a given genome (Nordborg and Weigel 2008; Ossowski et al. 2008; Huang et al. 2009; Arai-Kichise et al. 2011; Cao et al. 2011; Huang et al. 2013). With the increasing number of completely sequenced and assembled plant genomes (Feuillet et al. 2010) and especially in genomes with a relatively small genome ($1C = < 1$ pg or < 1 Gbp), this approach is rapidly becoming routine now. Meanwhile, it has been used for a number of additional plant species due to cost reductions in generating large amounts of sequence data. For example, for the tomato genome, the resequencing of genomes with a depth of 25X is now feasible for less than US \$ 5,000, and optimized bioinformatic procedures permit the identification of SNPs with $> 99\%$ accuracy. Currently, the genome resequencing approach is still cost-intensive but already feasible for large and highly complex plant genomes such as maize and soybean (Lai et al. 2010; Lam et al. 2010; Wu et al. 2010; Hufford et al. 2012; Jiao et al. 2012). A shortcut for a genome without a reference sequence is the use of assemblies of genes, including their intron and flanking sequences, allowing at least the identification of very large numbers of SNPs in gene sequences (You et al. 2011).

6.3 Selection of SNPs for a Genotyping Array

Once a large set of SNPs has been identified, several selection criteria need to be applied for the selection of high-quality SNPs to be placed on a genotyping array. If the SNPs along with their flanking sequences were collected from a variety of sources, the sequences need to be checked for overlaps between the data sets. Especially if SNPs are derived from data sets that were collected in the same germplasm pool and they result from different complexity reduction or transcriptome sequencing procedures, it is very likely that a certain number of SNPs will be present multiple times. Another point is that the SNP assay chemistry might put some constraints on the SNP itself or its flanking sequence (e.g. regarding the base composition for

an assay primer or oligonucleotide). That can eliminate a significant number of SNPs from being useful for assay design. Especially in species with a high level of polymorphism such as maize, potato or more general outbreeding species containing more than one SNP within 50 base pairs, it is very common that additional SNPs will be present adjacent to each other. In such a case, an SNP assay that does not consider this situation could result in a failed assay in a considerable proportion of the germplasm, depending on the allele frequency of the adjacent SNP. Thus, in essentially all SNP assay systems a certain region left and/or right should be devoid of additional SNPs to ensure the functionality of the assay in a wide range of genetic material. A peculiarity of the Illumina platform is that it requires two assay primers (features) for the analysis of some SNPs (A/T or G/C, Infinium I assays), but only one for other SNPs. This could significantly influence the number of analyzed SNPs that can be placed on an array. Finally, an optimal SNP selection for a genotyping array has to include the concept of haplotypes for each assayed locus. A haplotype is a set of SNPs that is in full linkage disequilibrium (LD) and can also be understood as an allele. Several SNPs that are in LD between two different haplotypes do not provide additional information, if they are assayed simultaneously (Pfeiffer and Gunderson 2009). This is illustrated in Fig. 6.1 where a number of SNPs in maize lines are in LD in each different haplotype. Selecting several SNPs for assay design that are diagnostic for the same specific haplotype does not provide additional information. Such SNPs should be eliminated during assay design and only SNPs should be selected that are discriminating between different haplotypes. This is especially important in species where the number of different haplotypes is as low as in tomato. In the 10 K array (Sim et al. 2012) it is relatively frequent that SNPs assayed in the same gene are derived from the same haplotype. Thus these SNPs are in full LD and do not provide additional information.

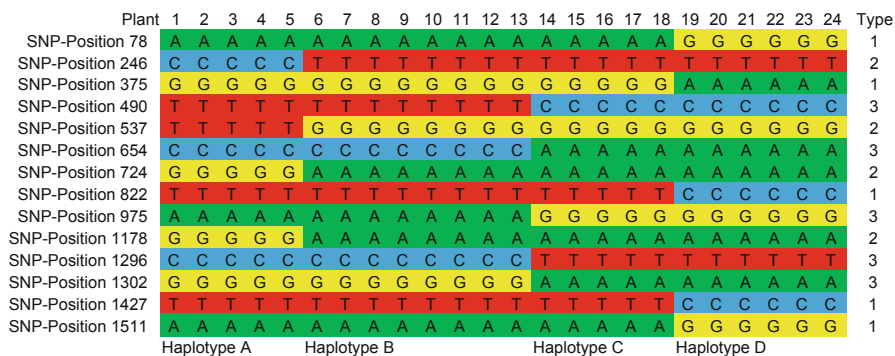


Fig. 6.1 Haplotype-based SNPs. The SNPs (in base codes) observed in a gene are displayed with their position in comparison to a reference sequence. Four different haplotypes (A, B, C, D) occur in this gene based on the analyzed individuals. All four haplotypes can be discriminated by three types of SNPs according to the allelic situation in the identified individuals. SNPs of the same type do not add additional information in this set of four haplotypes. Thus, only one SNP for each type is sufficient to collectively discriminate all haplotypes

While the selection of SNPs for an array in an unsequenced genome will most likely be based on polymorphisms in genes (Hasenmeyer et al. 2011), in a sequenced genome it is possible to use a number of different criteria for the selection of SNP markers for a large genotyping array. As in an unsequenced genome, SNPs could be selected that are predominantly located in genes, since genes and their flanking sequences are in most cases the basis of variation in a given organism. This approach carries the risk that genes that are expressed at a very low level will be missed in the SNP selection process and in some cases, such as rice, an alternative approach based on a more or less even physical distance between markers has been used (Zhao et al. 2011). Furthermore, SNP selection needs to take into account all knowledge on the level and type of genetic variation. For example, for maize it is known that the extent of LD in landraces and unadapted material is extremely low, while LD in commercial germplasm can extend over considerable distances (Lu et al. 2011; van Inghelandt et al. 2011). Thus, for such a genome it might be appropriate to select a hybrid approach, as it has been done by us in case of the maize genome. In this hybrid approach (Ganal et al. 2011) SNPs have been selected primarily to cover as many maize genes as possible with at least one or a few SNPs (17,520 genes with 33,417 SNPs) and these SNPs have been complemented by another set of SNPs (16,168 SNPs) not located within genes but instead distributed through the maize genome based on physical distance.

Going through all these above-mentioned criteria for maize and starting with set of more than 800,000 SNPs, this has resulted in a set of approximately 50,000 high quality SNPs that were distributed across the maize genome (Schnable et al. 2009) and which were finally put on the array. Summing up these experiences, it is advisable for the design of a high density SNP array to have at least 5–10 times the number of SNPs available which finally will be put on such an array.

6.4 SNP Calling Based on Array Data

Once an array has been designed and the first samples are being analyzed, it is important to assess the markers on the array for their quality with respect to calling the correct genotype in the investigated germplasm. During this process, a carefully selected set of lines has to be included. This set should contain lines and/or varieties that have been used in the SNP identification procedure as well as genetic material from the entire range of germplasm that will be analyzed with the array, and finally parent-F1 triplets from various crosses. Only with such material, a reliable allele calling can be established that can be used, for example, with the Illumina Infinium platform in routine genotyping via a so-called cluster file. A standardized procedure ensures that all investigated material is classified in the same way and the data from different laboratories and experiments can be merged into one data set.

SNP calling (Fig. 6.2) is usually straightforward in diploid species with a low level of polymorphism such as tomato or barley. A higher level of polymorphism can cause the failure of allele calling in a significant number of lines for a given marker due

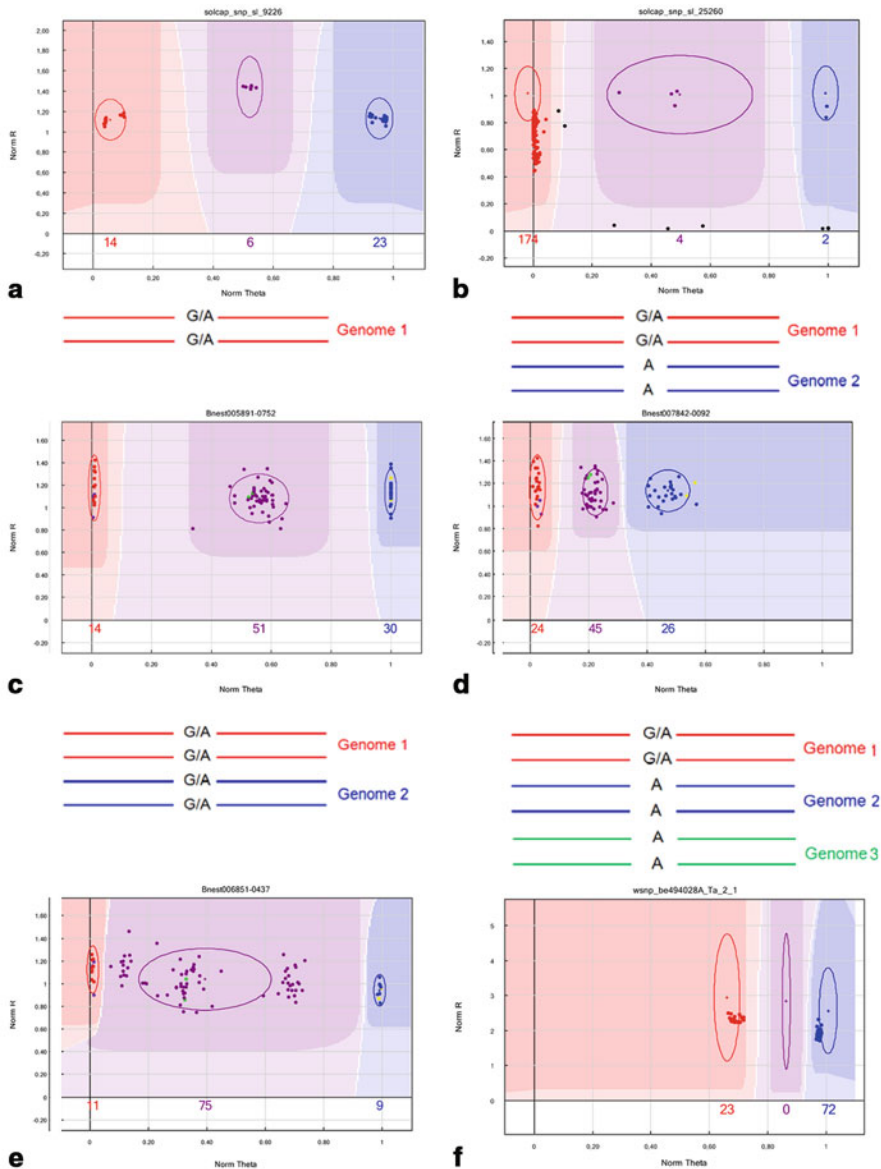


Fig. 6.2 SNP analysis in organisms with different ploidy levels (Illumina Infinium platform). **a** Typical pattern for a high quality SNP marker in a diploid organism. In the red area are individuals that are homozygous for allele 1, purple are heterozygous plants and blue are individuals that are homozygous for allele 2. **b** Typical pattern for an SNP marker in a diploid organism with an adjacent SNP not detected during assay design. Black dots at the bottom are individuals where the assay failed completely due to the flanking SNP influencing the functionality of the assay (null allele). In the red area are individuals that are homozygous for allele 1 (upper dots) or heterozygous for allele 1 and the null allele (lower dots), purple are heterozygous individuals for allele 1 and 2, and blue

to previously unknown adjacent polymorphisms in flanking sequences that hamper the functionality of the respective SNP assay in specific haplotypes (Ganal et al. 2011). Increasing levels of ploidy can complicate the allele calling, depending on whether it is an autopolyploid or an allopolyploid species. In an autotetraploid species such as potato, the respective locus can reflect five different allelic constitutions (AAAA, AAAB, AABB, ABBB and BBBB). This results in five different clusters which have to be clearly separated from each other for the correct allele calling. Only recently, this has been made possible for the Illumina Infinium platform. In an allopolyploid species such as oilseed rape or *Brassica napus* (Durstewitz et al. 2010) which consists of two different genomes (A and C) or wheat which consists of three different genomes (A, B and D), a large number of different patterns can occur when a specific SNP is analyzed (Akhunov et al. 2009; Chao et al. 2010). For example, in allotetraploid species up to three different cluster types can occur. If the respective SNP assay is functional in only one of the two genomes the pattern is identical to that of a diploid species with nicely distributed clusters for the three possible genotypes. If the SNP assay is functional in both genomes and thus assays both genomes, the clusters are shifted to one side in case only one of the two genomes is polymorphic since actually the allelic combinations of AAAA, AAAB and AABB are detected. This makes a cluster definition more difficult. If the SNP is present and segregating in both genomes at the same time, five clusters can be observed as in an autotetraploid species. Similar patterns are also observed with a considerable percentage of SNP assays in recently polyploidized genomes such as the maize genome. In the hexaploid wheat genome, even more different cluster types can be observed. If an SNP assay samples all three genomes but is polymorphic in only one of them, the actually analyzed allele combinations will be AAAAAA, AAAAAAB and AAAAABB. Two genomes will be monomorphic in the background resulting in very tight clusters that can only be separated with difficulty especially in heterozygous samples. Furthermore, depending on the number of assayed genomes, all other cluster patterns can occur that are observed in diploid and allotetraploid species (if only one

Fig. 6.2 (continued) are individuals that are homozygous for allele 2. **c** Typical pattern of a high quality SNP marker in an allotetraploid species (e.g. oilseed rape) where only one of the two genomes is detected. The other genome is either not present at this locus or the other genome is not analyzed due to polymorphism(s) between the two genomes that result in a failed assay in one of the genomes. The data interpretation is as for an SNP in a diploid species. **d** Typical pattern of a high quality SNP marker in an allotetraploid species (e.g. oilseed rape) where one genome is polymorphic and the other genome is monomorphic in the background. This generates three clearly defined clusters shifted to one side since the three clusters represent for the G/A polymorphism the situation GGGG, GGGA and GGAA. **e** Typical pattern of an SNP marker in an allotetraploid species where both genomes are polymorphic at the same time. This generates five clusters representing for the G/A polymorphism the situations GGGG, GGGA, GGAA, GAAA and AAAAA (three of them are in the heterozygous area when using the typical analysis software). For an autotetraploid species (e.g. potato), the five clusters can be called with a special software extension. **f** Typical pattern of a high quality SNP marker in an allohexaploid species (e.g. wheat) where one genome is polymorphic and the other two genomes are monomorphic in the background. This generates also three clearly defined clusters with the clusters heavily shifted to one side since the three clusters represent for the G/A polymorphism the situation GGGGGG, GGGGGA and GGGGAA

or two of the three genomes are assayed). SNPs that are segregating in more than one genome at the same time cannot be scored in wheat. An alternative, though currently not fully validated approach to solve these allele calling difficulties in allopolyploid species could be that the SNP assays are already selected and designed during the SNP identification and selection procedure in a way that the respective SNP assays are intentionally converted into genome-specific assays through the use of flanking SNPs that differentiate between the individual genomes so that the pattern complexity is reduced beforehand.

6.5 Analysis of SNP Genotyping Data from a Large Array

When a standardized allele calling procedure has been established based on a cluster file generated with a representative sample of lines or varieties, the genotype data for the thousands of SNPs can be used for a first data analysis and data assembly in order to determine the overall quality of the array. With respect to the technical quality of the array, this includes a number of features such as the reproducibility obtained for a number of DNA replicates which should essentially be 100 % and the consistency in parent-F1 triplets. High quality markers should be very consistent in generating the correct genotype calls and usually this value should be higher than 99.9 % in parent-F1 triplets. If this is not the case, it should be checked whether such markers have problems with allele calling and should be eliminated from future analyses. For markers that are reproducible and consistent in parent-F1 triplets, it might be appropriate to further define a certain percentage of failed samples above which a given SNP marker is eliminated from the analysis.

Once this initial technical data analysis regarding the general functionality of the SNP markers has been performed, a second data analysis and data assembly is dependent on the fact whether a reference genome is available or not. If a reference genome is available, scored markers should be ordered according to their physical mapping position in the genome and more specifically on the individual chromosomes. Such a data assembly will be of high value for the subsequent data analysis since the haplotype structure of the markers in the investigated material can be easily identified. If a reference genome sequence is not available, many SNP markers could possibly be ordered based on a draft genome assembly or if they are located within genes based on a virtual genome order as it is for example available for some species in the Gramineae (Mayer et al. 2011) or alternatively based on integrated genetic maps from a number of segregating populations.

6.6 Large SNP arrays in crop plants and examples for their use

In the first part of the following section, we provide an overview over the currently available large genotyping arrays for crop plants. In the second part, we show some examples from our own research regarding the use of such genotyping arrays with the main focus on genetic diversity analysis and the analysis of haplotype structure.

6.6.1 Availability of Large Genotyping Arrays for Crop Plants

In 2010 and 2011, first genotyping arrays with thousands of SNPs have been arising for some crop plants. In rye (*Secale cereale*), a 5 K array based on SNPs identified through transcriptome sequencing has been published (Hasenmeyer et al. 2011). It has been used for the characterization of rye germplasm and varieties. A 9 K SNP array has been assembled for grape (*Vitis* spec.). This array has been used for the characterization of a large number of lines from the genus *Vitis* as well as for the study of the genetic structure and domestication history of cultivated grape (Myles et al. 2010a, b). An array containing 44100 SNPs has been published for rice (*Oryza sativa*). This array has been used for the genotyping of rice accessions and the association of markers with 34 different traits (Zhao et al. 2011). The maize 50 K array contains 49585 SNP markers that can be scored over a wide range on maize germplasm. This array has been used for the analysis of several hundred maize lines and the generation of two genetic maps (Ganal et al. 2011). Finally, an 8K apple array (Chagnè et al. 2011) with 7867 apple SNPs was used for the analysis in segregating families and a germplasm collection.

Recently (mostly in 2012), a considerable number of arrays have been published for additional crop plants including sunflower (Bachlava et al. 2012), tomato (Sim et al. 2012), potato (Felcher et al. 2012), barley (Comadran et al. 2012), soybean (Song et al. 2013) and peach (Verde et al. 2012), which are all diploid species. For 2013 and 2014, it is expected that large genotyping arrays for the allopolyploid oilseed rape (*Brassica napus*) and hexaploid wheat (*Triticum aestivum*), which are more challenging in their accurate scoring, will be published.

6.6.2 Examples for the Use of Large Genotyping Arrays for the Characterization of Plant Germplasm and Varieties

Large genotyping arrays provide information about the genetic structure of individuals at a genome-wide resolution. With this information, it is possible to obtain new insights into the genetic constitution of specific germplasm types within a species. For example, in maize it has been reported that in general the extent of linkage disequilibrium (LD) is quite small. However, data from breeding material suggest that in such material LD could extend over much larger regions (Yan et al. 2009). With the 50K genotyping array, it is possible to study maize germplasm in detail. As shown for one region of the maize genome in Fig. 6.3, the extent of LD can vary greatly in different germplasm groups. In this way, novel insight about the genetic architecture and population structure can be gained that could have a deep impact on plant breeding and germplasm improvement.

At present, there is an increased interest in another type of genetic polymorphism which is different from simple sequence polymorphisms. Such polymorphisms are termed copy number variation or presence/absence variation (Swanson-Wagner

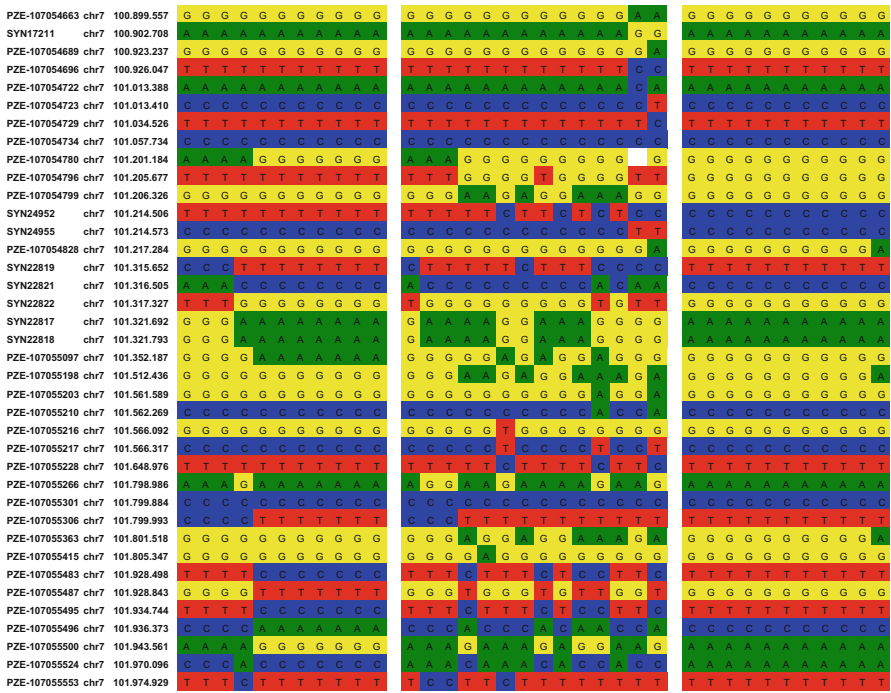


Fig. 6.3 Extent of linkage disequilibrium in different types of maize germplasm. Representative lines from three different types of maize inbred lines are shown for a region of chromosome 7. Markers, chromosomal assignment and physical position (in kb) of each marker are shown in the first three columns. Genotype data displayed as base calls in the left group represent samples from the female gene pool of a European hybrid maize program. Genotype data in the right group represent samples from the male gene pool of a European hybrid maize program. In the center are a set of representative inbred lines from cultivated temperate maize material containing European and North American founder lines. The three different germplasm pools show clearly different numbers of haplotypes and different levels of LD

et al. 2010). It is difficult to identify such polymorphisms by the use of genotyping arrays because the scoring of individual SNPs for presence or absence could also be influenced by the quality of the individual marker. However, presence/absence polymorphisms can be readily identified when the allelic situation in a specific region is analyzed based on the haplotype structure. For example, if a number of adjacent markers are either present or absent in specific lines including the occurrence of specific haplotypes, then it is very likely that in this regions a certain block of DNA sequences is either present or absent (Fig. 6.4).

Large genotyping arrays usually have a lower marker density around the centromeres. This is frequently due to the fact that the centromeric regions are mainly populated by large blocks of heterochromatin (Heslop-Harrison and Schwarzacher, 2011) mostly consisting of highly repetitive sequences which cannot be analyzed with genotyping arrays. A possible concern could be that due to the lower marker

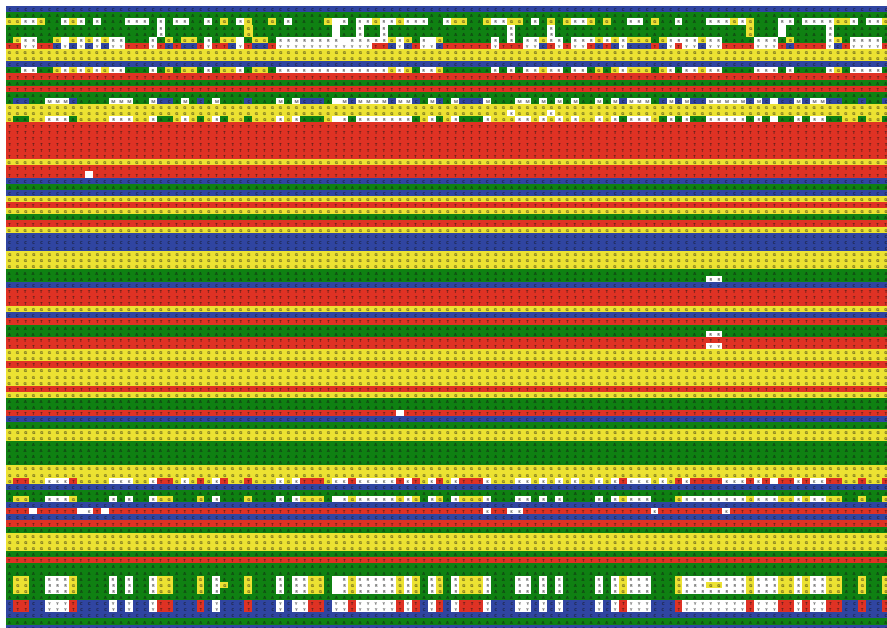


Fig. 6.5 Reduced level of polymorphism in the centromeric region of chromosome 10 of tomato. Genotype data are displayed for a number of tomato inbreds and hybrids representing the entire cultivated germplasm pool. Markers are ordered according to the tomato reference genome sequence. The upper and lower section displayed in the genotype data show a number of clearly defined haplotypes while the central section is represented by essentially only a single haplotype. The region with this single haplotype represents 37 Mio. bases or more than half of the entire physical length of the chromosome. On the genetic maps, this region is mostly cosegregating and devoid of recombination

6.7 Summary and Future Trends

Through the use of large scale genome sequencing technologies for transcriptome sequencing, reduced complexity sequencing or whole genome sequencing, the identification of essentially unlimited numbers of SNPs has become feasible. This will lead to the development of genotyping arrays containing many thousands of SNP markers for most or all important crop plants. Such genotyping arrays will permit the analysis of large numbers of lines and varieties with a standardized technology, resulting in genotyping data that can be compared directly between different experiments and laboratories. In genetic research and especially in plant breeding, this will permit the establishment of large databases with standardized genotyping data, and it will become possible to characterize genetic material from any plant species in much greater depth with respect to population structure, genetic relationships, loci involved in domestication, the extent of linkage disequilibrium and other aspects.

In plant breeding, large sets of genotyping data from arrays are currently opening the door to new breeding applications. These include the fast and precise mapping

of genes and QTLs. Furthermore, genotyping data from arrays with many thousands of markers permit the association of specific markers with specific traits. While association studies have been performed in human genetics for many quantitatively inherited traits, resulting in the identification of individual genes that have an effect onto such complex traits, such projects so far, have not been widely established in crop plants due to the lack of markers on the one hand and precisely phenotyped material on the other hand (Rafalski 2010). Even in the case where many loci with small effects control agronomically interesting traits, the use of large genotyping arrays permits the improvement of these traits through a process of Genomic Selection. This has already been validated in animal breeding and through first experiments in plants. Genomic Selection will certainly be a driver towards the development of increasingly larger arrays in the major crop plants in the future (Hamblin et al. 2011).

Currently, there is increasing discussion on the replacement of genotyping arrays through a process termed genotyping-by-sequencing involving the use of complexity reduction and bar-codes for the simultaneous sequencing of many individuals in combination with high throughput sequencing technologies (Davey et al. 2011; Elshire et al. 2011; Poland et al. 2012; Truong et al. 2012; van Poecke et al. 2013). While genotyping-by-sequencing has the potential to generate SNP data from larger numbers of loci than the currently used arrays (e.g. in maize), it is not clear to what extent this new technology can indeed replace array-based genotyping technologies. One issue is that genotyping-by-sequencing data so far lack of standardization. Genotype data generated in different projects can currently not easily be compared between different experiments and laboratories (especially in species without a high quality reference sequence) while array based data from different sources are easily comparable. Another issue is that genotyping-by-sequencing and array-based genotyping technologies compete in terms of cost which are highly dependent on the number of analyzed markers per line. Both sequencing costs per Mbp as well as costs per SNP on arrays are currently getting significantly lower and it is not easy to predict which technology will be cheaper in the near future. In the long term and given the efforts to reduce sequencing costs by orders of magnitude, it will however be clear that ultimately genome sequencing technologies will be applied for many genetic analyses (Huang et al. 2009, 2010, Schneeberger and Weigel 2011).

Acknowledgements The assistance of the technical staff at TraitGenetics during SNP marker development and the analysis of many samples using genotyping arrays is acknowledged. Large scale genotyping research at TraitGenetics has in part been funded by several grants from the German Federal Ministry of Education and Research (BMBF).

References

- Akhunov E, Nicolet C, Dvorak J (2009) Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *Theor Appl Genet* 119:507–517
- Arabidopsis Genome Initiative (2000) Analysis of the genome of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815

- Arai-Kichise Y, Shiwa Y, Nagasaki H et al (2011) Discovery of genome-wide DNA polymorphisms in a land race cultivar of Japonica rice by whole-genome sequencing. *Plant Cell Physiol* 52:274–282
- Bachlava E, Taylor CA, Tang S et al (2012) SNP discovery and development of a high-density genotyping array for sunflower. *PLoS One* 7:e29814
- Barbazuk WB, Emrich SJ, Chen HD et al (2007) SNP discovery via 454 transcriptome sequencing. *Plant J* 51:910–918
- Barchi L, Lanteri S, Portis E et al (2011) Identification of SNP and SSR markers in eggplant using RAD tag sequencing. *BMC Genomics* 12:304
- Blanca J, Esteras C, Ziarolo P et al (2012) Transcriptome sequencing for SNP discovery across *Cucumis melo*. *BMC Genomics* 24:280
- Chagné D, Crowhurst RN, Troggio M et al (2011) Genome-wide SNP detection, validation, and development of an 8 K SNP array for apple. *PLoS One* 7:e31745
- Cao J, Schneeberger K, Ossowski S et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 43:956–963
- Chao S, Dubcovsky J, Dvorak J et al (2010) Population and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). *BMC Genomics* 11:727
- Comadran J, Kilian B, Russell J et al (2012) Natural variation in a homolog of *Antirrhinum* CEN-TRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nat Genet* 44:1388–1392
- Davey JW, Hohenlohe PA, Etter PD et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev. Genet* 12:499–510
- Deschamps S, la Rota M, Ratashak JP et al (2010) Rapid genome-wide single nucleotide polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina genome analyzer. *Plant Genome* 3:53–68
- Durstewitz G, Polley A, Plieske J et al (2010) SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. *Genome* 53:948–956
- Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379
- Felcher KJ, Coombs JJ, Massa AN et al (2012) Integration of two diploid potato linkage maps with the potato genome sequence. *PLoS One* 7:e36347
- Feuillet C, Leach JE, Rogers J et al (2010) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16:77–88
- Fu Y, Springer NM, Gerhardt DJ et al (2010) Repeat subtraction-mediated sequence capture from a complex genome. *Plant J* 62:898–909
- Ganal MW, Altmann T, Röder MS (2009) SNP identification in crop plants. *Curr Opin Plant Biol* 12:211–217
- Ganal MW, Durstewitz G, Polley A et al (2011) A large maize (*Zea mays* L.) SNP genotyping array: Development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One* 6:e28334
- Gore MA, Chia J-M, Elshire RJ et al (2009a) A first-generation haplotype map of maize. *Science* 326:1115–1117
- Gore MA, Wright MH, Ersoz ES et al (2009b) Large-scale discovery of gene-enriched SNPs. *Plant Genome* 2:121–133
- Gunderson KL, Steemers FJ, Ren H et al (2006) Whole-genome genotyping. *Methods Enzymol* 410:359–76
- Hamblin MT, Buckler ES, Jannink JL (2011) Population genetics of genomics-based crop improvement methods. *Trends Genet* 27:98–106
- Han Y, Kang Y, Torres-Jerez I et al (2011) Genome-wide SNP discovery in tetraploid alfalfa using 454 sequencing and high resolution melting analysis. *BMC Genomics* 12:350
- Hasenmeyer G, Schmutzer T, Seidel M et al (2011) From RNA-seq to large-scale genotyping: genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol* 11:131

- Heslop-Harrison JS, Schwarzacher T (2011) Organisation of the plant genome in chromosomes. *Plant J* 66:18–33
- Huang X, Feng Q, Qian Q et al (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res* 19:1068–1076
- Huang X, Wei X, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in land races. *Nat Genet* 42:961–967
- Huang X, Lu T, Han B (2013) Resequencing rice genomes: an emerging new era of rice genomics. *Trends Genet*. doi: 10.1016/j.tig.2012.12.001
- Hufford MB, Xu X, van Heerwaarden J et al (2012) Comparative population genomics of maize domestication and improvement. *Nat Genet* 44:808–811
- Hyten DL, Cannon SB, Song Q et al (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* 11:38
- Imelfort M, Duran C, Batley J, Edwards D (2009) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotech J* 7:312–317
- Iorizzo M, Senalik DA, Grzebelus D et al (2011) De novo assembly and characterization of the carrot transcriptome reveals novel genes, new markers, and genetic diversity. *BMC Genomics* 12:389
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Jiao Y, Zhao H, Ren L et al (2012) Genome-wide genetic changes during modern breeding of maize. *Nat Genet* 44:812–815
- Lai J, Li R, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42:1027–1030
- Lam HM, Xu X, Liu X et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* 42:1053–1059
- Lu Y, Shah T, Hao Z et al (2011) Comparative SNP and haplotype analysis reveals a higher genetic diversity and rapider LD decay in tropical than temperate germplasm in maize. *PLoS One* 6:e24861
- Mayer KF, Martis M, Hedley PE et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 23:1249–1263
- McGall GH, Christians FC (2002) High-density genechip oligonucleotide probe arrays. *Adv Biochem Eng Biotechnol* 77:21–42
- Metzker ML (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Myles S, Boyko AR, Owens CL et al (2010a) Genetic structure and domestication history of the grape. *Proc Natl Acad Sci U S A* 108:3530–3535
- Myles S, Chia JM, Hirwitz B et al (2010b) Rapid genomic characterization of the genus *Vitis*. *PLoS One* 5:e8219
- Novaes E, Drost DR, Farmerie WG et al (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9:312
- Nordborg M, Weigel D (2008) Next-generation genetics in plants. *Nature* 456:720–723
- Ossowski S, Schneeberger K, Clark RM et al (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 18:2024–2033
- Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7:e32253
- Peiffer DA, Gunderson KL (2009) Design of tag SNP whole genome genotyping arrays. *Methods Mol Biol* 529:51–61
- Rafalski JA (2010) Association genetics in crop improvement. *Curr Opin Plant Biol* 13:174–180
- Saintenac C, Jiang D, Akhunov ED (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat. *Genome Biol* 12:R88

- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Schneeberger K, Weigel D (2011) Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci* 16:282–288
- Sim SC, Durstewitz G, Plieske J et al (2012) Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One* 7:e40563
- Song Q, Hyten DL, Jia G et al (2013) Development and evaluation of SoySNP50K, a high-density genotyping array for soybean. *PLoS One* 8:e54985
- Stemers FJ, Chang W, Lee G et al (2006) Whole-genome genotyping with the single-base extension assay. *Nat Methods* 3:31–33
- Swanson-Wagner RA, Eichten SR, Kumari S et al (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* 20:1689–1699
- Trebbi D, Maccaferri M, de Heer P et al (2011) High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.). *Theor Appl Genet* 123:555–569
- Trick M, Long Y, Meng J, Bancroft I (2009) Single nucleotide polymorphism (SNP) discovery in the polyploid *Brassica napus* using Solexa transcriptome sequencing. *Plant Biotechnol J* 7:334–346
- Truong HT, Ramos AM, Yalcin F et al (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. *PLoS One* 7:e37565
- Van Inghelandt D, Reif JC, Dhillon BS et al (2011) Extent and genome-wide distribution of linkage disequilibrium in commercial maize germplasm. *Theor Appl Genet* 123:11–20
- Van Orsouw NJ, Hogers RCJ, Janssen A et al (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS One* 2:e1172
- Van Poecke R, Maccaferri M, Tang J et al (2013) Sequence-based SNP genotyping in durum wheat. *Plant Biotech J* 11:809–817
- Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 27:522–530
- Verde I, Bassil N, Scalabrin S et al (2012) Development and evaluation of a 9K SNP array for peach by internationally coordinated SNP detection and validation in breeding germplasm. *PLoS One* 7:e35668
- Winfield MO, Wilkinson PA, Allen AM et al (2012) Targeted re-sequencing of the allohexaploid wheat exome. *Plant Biotechnol J* 10:733–742
- Wu X, Ren C, Joshi T et al (2010) SNP discovery by high-throughput sequencing in soybean. *BMC Genomics* 11:469
- Yan J, Shah T, Warburton ML et al (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS One* 24:e8451
- You FM, Huo N, Deal KR et al (2011) Annotation-based genome-wide SNP discovery in the large and complex *Aegilops tauschii* genome using next-generation sequencing without a reference genome sequence. *BMC Genomics* 12:59
- Zhao K, Tung CW, Eizenga GC et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* 2:467

Chapter 7

Paleogenomics as a Guide for Traits Improvement

Jérôme Salse

Contents

7.1	Introduction	132
7.1.1	Genome Sequences Available and Sequencing Strategies	133
7.1.2	Comparative Genomics Methods, Data and Online Tools	138
7.1.3	Plant Genome Ancestors and Reconstructed Karyotypes	142
7.1.4	CAR aNd Derived COS for Genetic and Physical Mapping	151
7.1.5	Complex Traits Dissection	158
7.2	Future ChalLenges	165
	References	165

Abstract Technological advances in sequencing methodologies has led to the development and access of large sets of plant genomic resources during the last decade which has enabled high resolution comparative genomic studies that can be considered as a reference framework for improved (i) evolutionary analyses, (ii) physical/genetical mapping initiatives, and (iii) complex trait dissection studies. In term of paleogenomics-based evolutionary data, synteny-based identification of seven shared duplications in plants led to the modelisation of common ancestral genomes structures of 30–50 Mb with 5 to 7 protochromosomes and containing 10,000 to 15,000 founder protogenes. These reconstructed ancestral genomes led to an improved representation of plant genome syntenies in concentric crop circles, providing a new reference tool as a guide for accurate and improved gene annotation, cross-genome markers development and translational genomics-based modern trait dissection.

Keywords Comprative genomics · Pelogenomics · Translational biology · Traits improvement

J. Salse (✉)

INRA, UMR 1095 ‘Généétique, Diversité et Ecophysiologie des Céréales’,
Laboratory ‘Plant Paleogenomics for Traits Improvement’,
Domaine de Crouelle, 234 avenue du Brézet,
63100 Clermont Ferrand, France
e-mail: jsalse@clermont.inra.fr

Abbreviations

AB-QTL	Advanced Backcross QTL
AK	Ancestral Karyotype
BAC	Bacterial Artificial Chromosome
CALP	Cumulative Alignment Length Percentage
CAR	Conserved Ancestral Region
CD	Cluster Ratio
Cfis	Chromosome Fission
Cfus	Chromosome Fusion
CIP	Cumulative Identity Percentage
CNV	Copy Number Variation
COS	Conserved Orthologous Set
DR	Density Ratio
EST	Expressed Sequence Tag
HIF	Heterozygous Inbred Family
HRM	Highly Recombinant Material
HSP	High Scoring Pair
MYA	Million Years Ago
NIL	Near Isogenic Line
NGS	Next Generation Sequencing
PAV	Presence Absence Variation
QTL	Quantitative Trait Loci
R	Resistance
RIL	Recombinant Inbred Line
SNP	Single Nucleotide Polymorphism
WGD	Whole Genome Duplication
WGS	Whole Genome Shotgun

7.1 Introduction

Increasing access to the sequence of plant genomes based on the recent development of Next Generation Sequencing (NGS) technologies allows for inter- as well as intra-specific comparative genomics to an unprecedented resolution. A large number of public bioinformatics tools are available to perform comparative genomics studies in plants and animals. However, most of the comparative genomics studies lack standards for the identification of robust orthologous relationships between genomes leading non-specialists to often over-interpret the results of large-scale comparative sequence analyses. Recently, improved parameters and tools have been established to define significant relationships between genomes, especially once obtained from short NGS-derived sequence reads. Such standards are necessary to (i) identify robust sets of orthologous gene pairs, (ii) derive complete sets of chromosome-to-chromosome relationships within and between genomes and (iii) model common paleoancestor genome structures. The accurate reconstruction of founder ancestral

Table 7.1 BAC vs WGS strategies—The table details the main differences (*first column*) in species, resource sequencing assembly, coverage, time and cost related to the two main plant genome sequencing strategies, i.e. BAC-by-BAC (*second column*) and whole genome shotgun (*third column*) approaches

	BAC-by-BAC strategy	Whole Genome Shotgun strategy
Species	Cloning required	No restriction
Resource	BAC library and/or physical map	No restriction
Sequencing	No restriction	Mixed approach not limited to short reads
Assembly	No restriction	Difficult in high repeated regions
Coverage	Reduced	High
Time	High	Reduced
Cost	High	Reduced

genomes derived from the comparison of the gene content and structure of modern species can now be considered as a guide for the identification and the precise understanding, beyond description, of evolutionary mechanisms (such as duplications, translocations, inversions, fusions), as well as robust implementation of applied research tools in: (i) monitoring synteny-based computed gene order in non-sequenced genomes, (ii) characterising of universal gene-based markers for physical/genetical mapping purposes and (iii) conducting synteny-based dissection of agronomically important traits.

7.1.1 Genome Sequences Available and Sequencing Strategies

Approaches and tools used in comparative genomics studies really depend on the genome sequence quality and accuracy, directly connected to the sequencing strategies and approaches used. We will detail in the next section the different genome sequencing strategies and approaches classically used to unravel the advantages and limits in performing comparative genomics analyses.

7.1.1.1 Genome Sequencing Strategies

Since the first plant genome sequenced in 2000 with *Arabidopsis thaliana*, up to 20 sequenced flowering plant genomes have been released in the public domain. Two main strategies have been used, i.e. whole genome shot-gun (or WGS) and BAC-by-BAC (so called physical map-based or clones-based) approaches. In the BAC-by-BAC approach, a crude physical map needs to be constructed of the whole genome, chromosome or even locus prior sequencing. Constructing a map requires cutting the chromosomes into large (150 Kbp long) fragments (i.e. BAC library construction) of DNA and figuring out the order of these DNA segments (i.e. fingerprint) before sequencing all the fragments of the minimum tilling path, i.e. MTP (lowest number of BACs covering the entire considered region), Table 7.1. Each BAC of

Table 7.2 NGS strategies and approaches—Technical characteristics (*first column*) in strategy, amplification, platform, run length, run yield, read length and cost, related to the four main sequencing methods, i.e. Sanger (*second column*), 454 (*third column*), Solexa (*fourth column*) and SOLID (*fifth column*) approaches

Company	ABI	Roche			Illumina			ABI
Technology	Sanger	454			Solexa			Solid
Strategy	Polymerase	Pyrosequencing			Polymerase			Ligation
Amplification	No	emPCR			Solid-phase			emPCR
Platform	Capillary sequencer	GS20	FLX	Ti	GAI	GAI		#2
Run time ^a		6Hr	7Hr	9Hr	3D	3D	4D	5D
Run yield ^a	0.08	50Mb	100Mb	400Mb	1Gb	4Gb	6Gb	4Gb
Read length ^a	800	100	200	350	35	50	75	35
Cost (\$/Mbp)	5000	85	40	20	6	3	0.5	< 0.5

^a D for days, Hr for hours, Gb for Gigabases, Mb for Megabases, Kb for Kilobases

the MTP is shotgun sequenced, where many short reads are assembled to produce the sequence of the considered BAC clones. Overall, the whole genome sequence may be then obtained from these large assembled sequenced regions, based on the exploitation of BAC sequence overlap identity defining contiguous clones/contigs. Arabidopsis and rice genome sequences have been completed based on this previous clones-based approach in early 2000. International sequence centers then adopted a complementary strategy developed from the sequencing of vertebrates including the WGS strategy initially proposed by Weber and Myers in 1997 and then applied by Celera company. The WGS method goes directly in sequencing the genome randomly shredded into short fragments a few Kbp long, bypassing the need for BAC library/fingerprint/physical map construction steps. Essentially, the WGS strategy relies on making several insert size libraries, which are then sequenced from both ends (paired-end or mate-ends). Therefore, it is (i) applicable to all species including the ones known as refractory to cloning, (ii) much faster and, (iii) less expensive. However, sequence assembly and anchoring to chromosomes are more complex in the WGS approach, compared to the BAC-by-BAC strategy. The latter has provided a much more accurate and complete chromosome anchored sequence information (Table 7.1). Overall, where the first plant genome sequences (Arabidopsis and rice) have been achieved based on the BAC-by-BAC approach, the vast majority of recent sequences have been obtained following the NGS strategy. As a summary, WGS offers the access to the genomic sequence from one organism at once, where the clone-based approach focus on assembled and mapped contigs of BAC clones.

Following the burst in genome sequencing after 2009, recent advances in next generation sequencing (NGS) technologies offered additional opportunities to sequence (or re-sequence) the genomes of different plant species and lineages (for review Delseny et al. 2010). Three main NGS technologies are classically used. The Table 7.2 summarizes the advantages and limits of these complementary approaches. The first sequenced genomes have been achieved following the Sanger

method (briefly, amplification by cloning, primer extension with blocked and labelled nucleotides, separation by electrophoresis prior sequence reading), considered as an expensive strategy but very accurate with a low error rate compared to recent NGS technologies. The NGS differs from Sanger on the DNA fragment amplification and sequencing approaches. In contrast to Sanger technology, the pyrosequencing strategy (Ronaghi et al. 1998) also called sequencing by synthesis (i.e. during DNA extension) relies on reading the signal from the nucleotide during its incorporation. The most used pyrosequencing-based NGS approaches (i.e. 454, Solexa, SOLID) yield a large amount of sequence per run and are cost effective, Table 7.2 (for review Kircher and Kelso 2010). The Roche 454 consists of small DNA fragments ligated to adaptators and then separated into single strands that bound to small DNA capture beads. The fragment is then amplified upon ‘emulsion-PCR’, each beads carrying million copies of the DNA fragments suitable for pyrosequencing. The cloning approach used in the Sanger strategy is thus avoided. The Solexa sequencing system is based on small DNA fragments amplified on a solid substrate upon ‘bridge-PCR’. The SOLID approach relies on a new amplification (rolling circle) and sequencing (sequencing-by-ligation) strategies. More recent NGS technologies (i.e. Polonator, Helicos, Pacific) relying on single DNA molecule sequencing strategies are also available although they may be associated with much higher error rate compared to the previous popular NGS chemistries.

In summary, the advantages associated from NGS rely on avoiding cloning, low reaction volume cost and high speed, where the limitations consists in read length, library preparation (fragmentation/ligation/amplification) introducing bias and artefact based on contrasted error rates. NGS approaches need a gold standard, by comparing for a given species or at least for several loci, the coverage necessary to consider in using a mixed 454/Solexa sequencing approach to reach the error rate classically observed in Sanger sequencing. Standards may also be needed regarding the sequence assembly approaches and associated software. The two most popular sequence assembly tools are Newbler delivered with the Roche technology and Velvet adapted to the Illumina data.

The most recent plant genome sequence projects rely on a hybrid approach, including long and short reads from pyrosequencing strategies on long-insert fosmid or BAC-end sequences combined with contigs assembled from paired-end short reads. The access to paired-end data allows for the assembly of short read contigs into large scaffolds based on the characterisation of neighbour contigs associated with face-to-face, paired-end sequences in two distinct and initially separated contigs. It is expected that such whole genome sequence (including un-anchored ones) will produce genome references of reduced accuracy and poorer completeness compared to the initial rice and Arabidopsis genomes. Overall, once conducting any comparative genomics study, sequencing approaches (WGS vs. BAC-by-BAC) and associated sequencing technologies (long vs. short reads, pyrosequencing vs. Sanger) need to be considered with caution (upon read length, error rates, sequence coverage) in order to use the most appropriate synteny tools and criteria described in details in the next sections.

7.1.1.2 Released Plant Genome Sequences

The sequence of the two first plant genomes delivered, *Arabidopsis* and rice, were obtained following the classical clone-based approach (BAC library, anchored physical map, MTP selection) sequenced using Sanger technology (applied on sub-cloning into shotgun libraries of the MTP BAC clones). Pure NGS strategies (cucumber for example) and hybrid ones (Sanger + pyrosequencing) applied on WGS or BAC resources derived in a tremendous explosion in genome sequence release since 2009. Up to 20 flowering plant genome sequences are now available (*Arabidopsis*, AGI 2000; papaya, Ming et al. 2008; soybean, Schmutz et al. 2010; lotus, Sato et al. 2008; apple, Velasco et al. 2010; grape, Jaillon et al. 2007; cacao, Argout et al. 2011; strawberry, Shulaev et al. 2011; poplar, Tuskan et al. 2006; rice, IRGSP 2005; sorghum, Paterson et al. 2009; maize, Schnable et al. 2009 and *Brachypodium* IBI 2010) and the number of high-resolution gene based genetic maps (*Triticeae* for example, Qi et al. 2004 and Stein et al. 2007) is also increasing, both allowing to perform evolutionary comparative genomics and derived paleogenomics studies to an unprecedented resolution.

The quality of the genome drafts is heterogeneous in term of genome coverage, number of gaps and percentage of anchored genome sequence. When considering the sequence genome gold standards, consisting of a completely finished sequence with no (or few centromeric) gaps and with less than one error in 100 kb, then among the 13 genome drafts anchored to the chromosomes and reported in Table 7.3, only the rice and *Arabidopsis* genomes are close to this goal. Table 7.3 provides the list of genome sequences available as well as associated features such as genome size (from 0.1 up to 17 Gb), number of chromosome (from 5 to 21), gene/TE content (from 15691 to 58984), as well as the sequencing strategy used (as detailed in the previous section). Crop genome sequences reveal a direct relationship between the variation in genome size and the content in repeated sequences. The plant genomes differ largely in their organisation, i.e. chromosome number, physical size and polyploidy level as illustrated in Fig. 7.1 The differences in genome size are mainly correlated to the repetitive DNA content (Bennetzen 2005). A four-fold size change is reported in Table 7.3 between sorghum (659 Mb) and maize (2365 Mb) genomes that can be entirely explained by the difference in TE content (respectively 62% vs 84%), while the gene content is almost similar (respectively 34008 vs 32540). Rapid amplification and turnover of few TE families is sufficient for genome size differentiation between species as well as between relatives (Piegu et al. 2006) leading to erosion in intra-specific genome colinearity as reported initially between maize varieties (Fu and Donner 2002).

Despite differences in organisation, the chromosomal architectures present some similarities with gene rich telomeric and sub-telomeric regions in contrast to TE rich centromeric and peri-centromeric ones. Among the grass genomes that were and still are of high priorities for sequencing, NGS approaches open the way for decoding in the next future genomes of large (high repeat content such as wheat) complex (polyploids such as wheat, peanut and coffee) organisation. Numerous plant species are under sequencing and several expected to be release in 2012–2013 including

Table 7.3 Plant genome data sets used in paleogenomics studies (Adapted from Salse 2012)—The table details the 15 plant genome sequences available (*first column*) regarding the common name (*second column*), the sequencing strategy (either WGS or BAC-by-BAC, *third column*), number of chromosomes (*fourth column*), genome organisation (size in Mbp and TE content in % of genome coverage, *fifth column*) number of annotated genes (*sixth column*), synteny data (number of orthologs, number of blocks, % of genome covered, *seventh column*), duplication data (number of paralog, number of blocks, % of genome covered, *eighth column*), number of polyploidization events (R for rounds, *ninth column*)

Species	Name	WGS (W) or BAC (B)	Chr	Size (Mbp)	TE (%)	Annotated genes	Synteny	Duplication	WGD
<i>Eudicot</i>									
<i>Vitis vinifera</i>	Grape	W	19	302/22		21189	RG ^a	543-23-71	IR
<i>Arabidopsis thaliana</i>	Cress	B	5	119/19		33198	2389-80-99	1630-5-83	3R
<i>Populus trichocarpa</i>	Poplar	W	19	294/35		30260	4555-87-92	4164-46-73	2R
<i>Glycine max</i>	Soybean	W	20	949/50		46194	4013-164-97	9533-89-55	3R
<i>Carica papaya</i>	Papaya	W	9	234/31		19060	3199-65-75	215-36-55	IR
<i>Fragaria</i>	Strawberry	W	7	208/24		32630	3289-94-70	114-27-19	IR
<i>Theobroma cacao</i>	Cacao	W	10	218/26		27814	4472-21-81	370-19-66	IR
<i>Lotus japonicus</i>	Lotus	B	6	462/40		15691	1720-80-61	145-32-35	2R
<i>Malus domestica</i>	Apple	W	17	528/42		58984	3498-104-70	2845-69-59	2R
						Total	27135-695-81	19559-396-57	
<i>Monocot</i>									
<i>Oryza sativa</i>	Rice	B/W	12	372/39		41046	RG ^a	448-10-73	IR
<i>Sorghum bicolor</i>	Sorghum	W	10	659/62		34008	6147-12-99	409-10-84	IR
<i>Zea mays</i>	Maize	B	10	2365/84		32540	4454-30-82	3454-17-99	2R
<i>Brachypodium distachyon</i>	Brachypodium	W	5	271/28		27601	8533-12-99	642-13-79	IR
^a <i>Triticum aestivum</i>	Wheat	–	21	~ 17 Gb/> 80		5003 ^b	827-13-91	102-10-75	IR
^a <i>Hordeum vulgare</i>	Barley	–	7	~ 5000/> 80		3423 ^b	309-13-84	38-9-75	IR
						Total	20270-80-91	5093-69-81	

^a RG refers to Reference Genome indicating that rice (*Oryza sativa*) and grape (*Vitis vinifera*) have been used as reference genomes for the synteny analysis for the monocots and eudicots, respectively

^b Mapped genes or genic markers

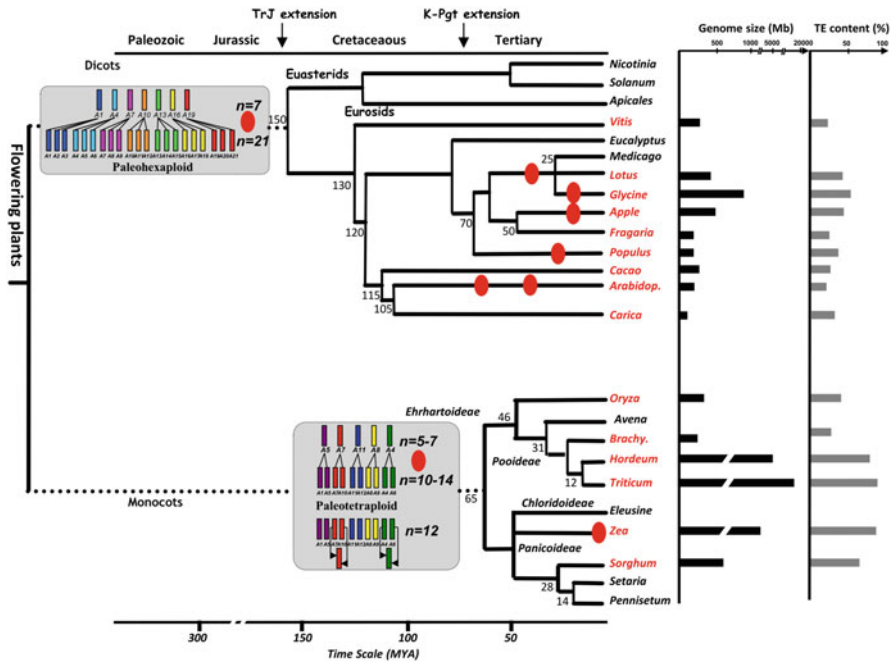


Fig. 7.1 Plant evolutionary tree (adapted from Abrouk et al. 2010)—Schematic representation of the phylogenetic relationships between plant species, i.e. eudicots (*top*) and monocots (*bottom*). Divergence times from a common ancestor are indicated on the branches of the phylogenetic tree and as the underneath scale (in millions years). Sequenced genomes are indicated in red. Whole genome duplication events are illustrated with red circles on the tree branches. The pre- ($n = 5$ and $n = 7$, for the monocots and eudicots) and post- ($n = 12$ and $n = 21$, respectively, for the monocots and eudicots) polyploidy ancestral karyotypes are indicated at the left of the tree. Finally, distribution of the genome sizes (as black bars expressed in Mb) and TE content (as grey bars expressed in % of genome coverage) are shown at the right side of the figure

Medicago truncatula, tomato, potato, millet, cotton, cassava, eucalyptus, and peach. We can speculate that most of the species of agronomical interest or cultivated ones will have their genome sequenced in the coming years.

7.1.2 Comparative Genomics Methods, Data and Online Tools

A large number of bioinformatic tools and approaches are now available to perform comparative genomics studies. However, most of the publicly available and, user-friendly tools and methods lack standards in the identification of robust orthologous relationships between genomes leading to often over- or mis-interpret the results of large-scale comparative sequence analyses. It becomes of primary importance to establish a number of improved parameters and tools to define significant relationships between the genomes in (i) identifying robust sets of orthologous gene pairs, and (ii) deriving complete sets of chromosome-to-chromosome relationships.

7.1.2.1 Comparative Genomics Parameters and Standards

Several ‘user friendly’ graphical displays of colinearity between plant genomes that will be presented in the next section are available to use and exploit comparative analyses derived from genome sequences alignment. However, the most important and critical step before visualizing alignments is to apply methods that will enable robust assessment of the relationships between the aligned sequences thereby ensuring relevant downstream interpretation (i.e. evolutionary inference and ancestor reconstruction) or application (i.e. marker development, physical/genetical mapping or trait dissection). Because it is difficult to infer orthologous (derived from a common ancestor by speciation) and paralogous (derived by duplication within one genome) relationships from sequence comparisons, stringent alignment criteria and statistical validation are now necessary to accurately evaluate whether the association between two or more genes found in the same order on two chromosomal segments occurs by chance or truly reflects colinearity, *cf* Fig. 7.2.

In order to precisely address gene conservation and duplication relationships, three sequence alignment parameters were recently defined in Salse et al. 2009a. The first parameter, AL (Aligned length), corresponds to the sum of all HSP lengths. The second, CIP (cumulative identity percentage) corresponds to the cumulative percent of sequence identity obtained for all the HSPs ($CIP = \sum nb \text{ ID by HSP} / AL \times 100$). The third parameter, CALP is the cumulative alignment length percentage. It represents the sum of the HSP lengths (AL) for all the HSPs divided by the length of the query sequence ($CALP = AL / \text{Query length}$). In summary, the CIP and CALP parameters allow the identification of the best alignment, i.e. the highest cumulative percentage of identity in the longest cumulative length, taking into account all HSPs obtained for any pairwise alignment, *cf* Fig.7.2a.

Regarding the validation of conserved and duplicated blocks, two criteria based on the complete set of either two paralogous or orthologous regions were recently defined (Salse et al. 2009a). These two related regions are characterised by a physical size (Size), a number of annotated gene (Gnumber), a number of orthologous or paralogous couples (Cnumber). The two derived criteria are the Density Ratio (DR) and the Cluster Ratio (CR): $DR = [(Size\ 1 + Size\ 2) / (2 \times Cnumber)] \times 100$ and $CR = [(2 \times Cnumber) / Gnumber\ 1 + Gnumber\ 2] \times 100$. The Density Ratio (DR) considers the number of links between two regions (duplicated or syntenic) in regard to the size of the considered blocks. The Cluster Ratio (CR) considers the number of links between two regions (duplicated or syntenic) in regard to the number of annotated genes available in the considered blocks. In summary, CR and DR allow the selection of the orthologous and paralogous block pairs associated with the highest number of respectively conserved and duplicated genes, *cf* Fig. 7.2b. Other similar approaches are available to identify conserved regions between and within genomes on a statistical basis: LineUp and derived CloseUp (Hampson et al. 2003) or ADHORE (Vandepoele et al. 2002) tools, for example.

Overall, large scale comparative genomics studies in plants including up to 20 distinct genomes require the use of accurate filtering standards that need to be established to ensure that different studies can be correlated and compared between

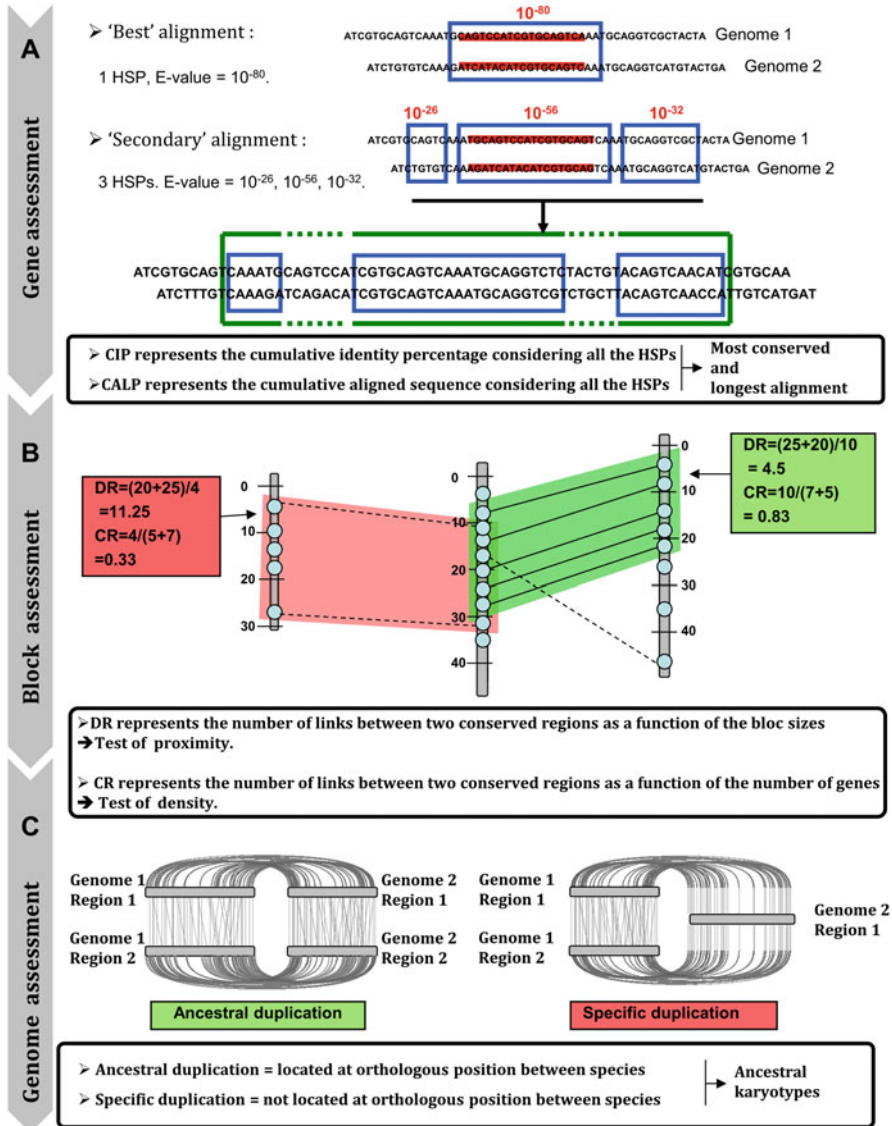


Fig. 7.2 Major principles of paleogenomics assessment between genomes—The figure illustrates precise examples of criteria used in addressing orthologous and paralogous gene relationships using CIP and CALP parameters (a), conserved and duplicated chromosomal blocks using CR and DR parameters (b), ancestral and specific duplications (c)

different research groups. The approach previously described to identify statistically-based gene pairs (either orthologs, paralog, homologs, homoeologs or ohnologs) can serve as a basis for developing further community standards in comparative studies at the inter- or intra-specific levels.

Table 7.4 Comparative genomics viewers—The table details the principal online comparative genomics viewers (*first column*) regarding the number of available species in 2012 (*second column*), the genome browser availability (*G. browser in third column*), the web address (*fourth column*) and associated references (*fifth column*)

Name	Nbr Species	G. browser	Web Address	Reference
NARCISSE	34	Yes	http://narcisse.toulouse.inra.fr/plants	Courcelle et al. 2008
PGDD	19	No	http://chibba.agtec.uga.edu/duplication/	Tang et al. 2008
Phytozome	19	No	http://www.phytozome.net/	none
CoGe	14	Yes	http://synteny.cnr.berkeley.edu/CoGe/	Lyons et al. 2008
GreenPhyl	11	No	http://greenphyl.cirad.fr/v2/cgi-bin/index.cgi	Rouard et al. 2011
PLAZA	9	Yes	http://bioinformatics.psb.ugent.be/plaza/	Proost et al. 2009
SynBrowse	3	Yes	http://www.synbrowse.org/	Pan et al. 2005
Cinteny	2	No	http://cinteny.cchmc.org/	Sinha and Meller 2007
InParanoid	2	Yes	http://inparanoid.sbc.su.se/cgi-bin/index.cgi	Ostlund et al. 2010
OrthologID	2	No	http://nypg.bio.nyu.edu/orthologid/	Egan et al. 2009

7.1.2.2 Plant Synteny Viewer Tools

Numerous tools have been developed and are now publicly available on the web to compare plant genomes and tentatively identify orthologs: NARCISSE, Plant Genome Duplication Database, Phytozome, CoGe, GreenPhyll, PLAZA, SynBrowse, Cinteny, Inparanoid, OrtholoID (Table 7.4).

These ‘user friendly’ tools differ in the number of species available (from 2 up to 34) and the graphical display. More importantly, such tools may sometimes rely on data obtained with low stringency alignment criteria and performed without systematic statistical validation as detailed in the previous section. In addition, they may not take into account the density and location of conserved genes to identify precisely paralogous and orthologous regions or segments and therefore overestimate colinearity between different segments of the genomes (both in term of number of conserved genes as well as regions/segments).

To address these issues, an online user friendly interface to access our comparative analyses, called ‘PlantSynteny’ has been recently proposed (Fig. 7.3a), that allows one to visualize orthologs as well as paralogs (Fig. 7.3b) among plant genomes either at the locus or gene levels, based on the alignment and validation standards described in the previous section. This web site provides access to the raw data (gene name, sequence, position and alignment criteria) obtained from the synteny and duplication analyses for the monocots (rice, maize, sorghum, wheat and barley) as well as for the eudicots (*Arabidopsis*, papaya, soybean, grape, poplar) genomes.

The image shows the 'GnpIS GENETIC AND GENOMIC INFORMATION SYSTEM' interface. The top header includes the logo and 'FEEDBACK'. The main title is 'Synteny / Plant Synteny Viewer'. On the left, there is a navigation menu with 'Parameter setting' (A) and 'Chromosome Synteny' (B) sections. The 'Parameter setting' section includes search parameters (gene name: Sb03g025240, ancestral chromosome: A5, modern chromosome: Select organum...) and display parameters (complementary search: Disable, modern chromosome window size: 41, ancestral chromosome window size: 31). The 'Chromosome Synteny' section shows a visualization of chromosomes for various species: Oryza sativa 1 and 5, Brachypodium distachyon 2, Sorghum bicolor 3 and 4, and Zea mays L. 3, 5, 6, and 8. Lines connect genes across chromosomes, with arrows indicating 'Paralogs' and 'Orthologs'.

Fig. 7.3 ‘PlantSynteny’ viewer tool (<http://urgi.versailles.inra.fr/synteny-cereal>)—The ‘PlantSynteny’ output layer displays the successive steps in parameter setting (a), chromosome relationship visualisation (b)

7.1.3 Plant Genome Ancestors and Reconstructed Karyotypes

In the past ten years, international initiatives have led to the development of large set of genomic resources and genome sequence comparison tools and approaches presented in the previous sections that allow comparative genomic studies between plant genomes at a high level of resolution. Comparison of genomic sequences revealed ancestral shared and lineage specific inter- and intra-genomic duplications, providing new insights into the evolution of flowering plant genomes from common ancestors. Plant genome synteny can be presented as concentric crop circles, providing a new reference for plant chromosome evolutionary relationships from ancestral karyotypes of reconstructed chromosome number, gene content and physical size.

7.1.3.1 Plant Genome Syntenies

Comparative genomics inference have first been obtained in the early 80's based on the comparison of genetic maps (mainly RFLP-based, i.e. Restriction Fragment Length Polymorphism) anchored with common molecular marker sets (so called map-based syntenies) concluding to an overall good level of gene conservation between plant genomes (for review Salse and Feuillet 2007). These initial comparative genomics studies were set on the use of homologous probes that cross-hybridize (or amplify) between species being compared. Overall, these results were obtained from low resolution genetic maps with an average of one marker every 10 cM that allowed the detection of only macro-scale chromosome-to-chromosome relationships as well as large rearrangements. Moreover, the maps were constructed with low copy molecular markers that were selected for their ability to provide a signal in cross hybridizations/PCR, thereby limiting the detection of whole or partial genome rearrangement events. It was also difficult to assess orthologous and paralogous relationships in large gene families since comparative mapping by RFLP often identified homologs rather than strict orthologous relationships, leading to a bias towards the identification of colinear regions.

Nowadays, more than 10 flowering plant genome sequences are available (*Arabidopsis*, papaya, soybean, lotus, apple, grape, cacao, strawberry, poplar, rice, sorghum, maize and *Brachypodium*, as referenced in the Table 7.3) and the number of high-resolution gene based genetic maps (*Triticeae* for example, referenced in the Table 7.3) is also increasing, both resources allowing evolutionary comparative genomics studies to an unprecedented resolution. Paleogenomics, or reconstruction of the ancestral genome structure of modern species, is based on large-scale comparative genomic analyses to identify accurately shared chromosomal structures and shuffling events. Using newly defined sequence alignment parameters and statistical validation criteria described the previous Section (6.2.2) to compare plant genomes that diverged from a common ancestor 150–300 million years ago (hereafter mya), 27135 orthologous genes defining 685 colinear blocks and covering on average 81 % of the eudicot genomes and 20270 orthologous genes defining 80 colinear blocks and covering on average 91 % of the monocot genomes, have been precisely characterised, *cf* Table 7.3 (Salse 2012). These data support the conclusion that between 10–20 % (for more than 50 mya of divergence exemplified by rice-sorghum speciation with 14 %) and 60–80 % (for less than 20 mya of divergence exemplified by sorghum-maize speciation with 73 %) of the genes are conserved as strict orthologs when (i) sequence conservation and maintenance, (ii) gene order and (iii) orientation, are taken into consideration in comparing genomes (*cf* Table 7.3).

The syntenic relationships observed between plant genomes have been classically illustrated through a pioneering model of circular consensus genetic maps of grasses, the so-called crop circles initiated by Mike Gale and co-workers (Moore et al. 1995; Devos and Gale 1997, 2000; Devos 2005; Gale and Devos 1998), where the genomes were arranged as concentric circles according to their genome size. The crop circles have recently re-updated from sequence genome based comparative genomics data described previously in this section (Bolot et al. 2009; Abrouk

et al. 2010). The crop circles in the Fig. 7.4 illustrates clearly the chromosome-to-chromosome conservation (grey lines between circles as orthologous genes) observed in the monocot (exemplified by the cereal circle as Fig. 7.4a involving *Brachypodium*/rice/sorghum/maize/Triticeae) and the eudicot (exemplified by the legume circle as Fig. 7.4b including Lotus/Medicago/soybean/pea), Abrouk et al. 2010; Salse and Feuillet 2011; Bordat et al. 2011; Murat et al. 2012; Salse 2012).

Thus, based on this representation of the chromosome-to-chromosome conserved synteny relationships (illustrated with a colour code that illuminate the ancestral karyotype structures described in the next Section 6.3.3), it is possible to immediately identify for any radius of the ‘crop circles’ the ancestral relationships and origins (WGD, breakages, chromosome fusions) of the different chromosomes in each of the nine modern cereal and legume genomes (Fig. 7.4). For example, one of the ancestral duplication (between A1 and A5 illustrated as purple dark/light blocks) involved modern chromosomes 1-5, 3-6-8, 3-9, 1-3 respectively in rice, maize, sorghum and the Triticeae. Despite the global conservation of gene content and order between plant genomes, intergenic regions have been subject to different rate of repeat sequence invasion (see for example microcolinearity example in Fig. 7.4a for the cereal and 4B for the legume circles) so that no orthologous sequences are found in such non-genic sequences. Overall, it is possible to visualize the ancestral relationships and the origin of the chromosome of modern plant genomes at both genome-wide and locus levels (Salse and Feuillet 2011; Salse 2012).

7.1.3.2 Plant Genome Duplications

Modern diploid species have long been proposed as ancestral polyploids, since the early 1980s, based on the comparison of genetic maps anchored with common markers (for review Salse and Feuillet 2008 as well as Vandepoele et al. 2003; Van de Peer 2009a; Jiao et al. 2011). Nowadays, based on the plant genome sequences available (Table 7.3) associated with the sequence comparison methodologies described in the previous Section (6.2.1), all eudicot genomes although diploid in their current structure, contain duplicated genes (19559 paralogous gene pairs in total) defining paralogous segments (396 duplicated blocks in total) that cover in average 57 % of the considered genomes (Table 7.3). In the same way, 5093 paralogous gene pairs in total defining 69 paralogous segments total covering in average 81 % of the considered monocot genomes has been characterised. The crop circles in Fig. 7.4 clearly illustrate the duplicated nature (dark vs. light coloured blocks within circles) observed in the monocot (exemplified by the cereal circle in Fig. 7.4a) and the eudicot (exemplified by the legume circle in Fig. 7.4b).

Integration of intra-species duplication and inter-species synteny analyses in the eudicot and monocot genomes allowed for a precise characterisation of seven shared ancestral duplications recovered in all the investigated plant genomes. Such duplications cover more than 50 % of any considered genome in eudicots and monocots, a clear proof of whole genome duplication (WGD or R) events, demonstrating

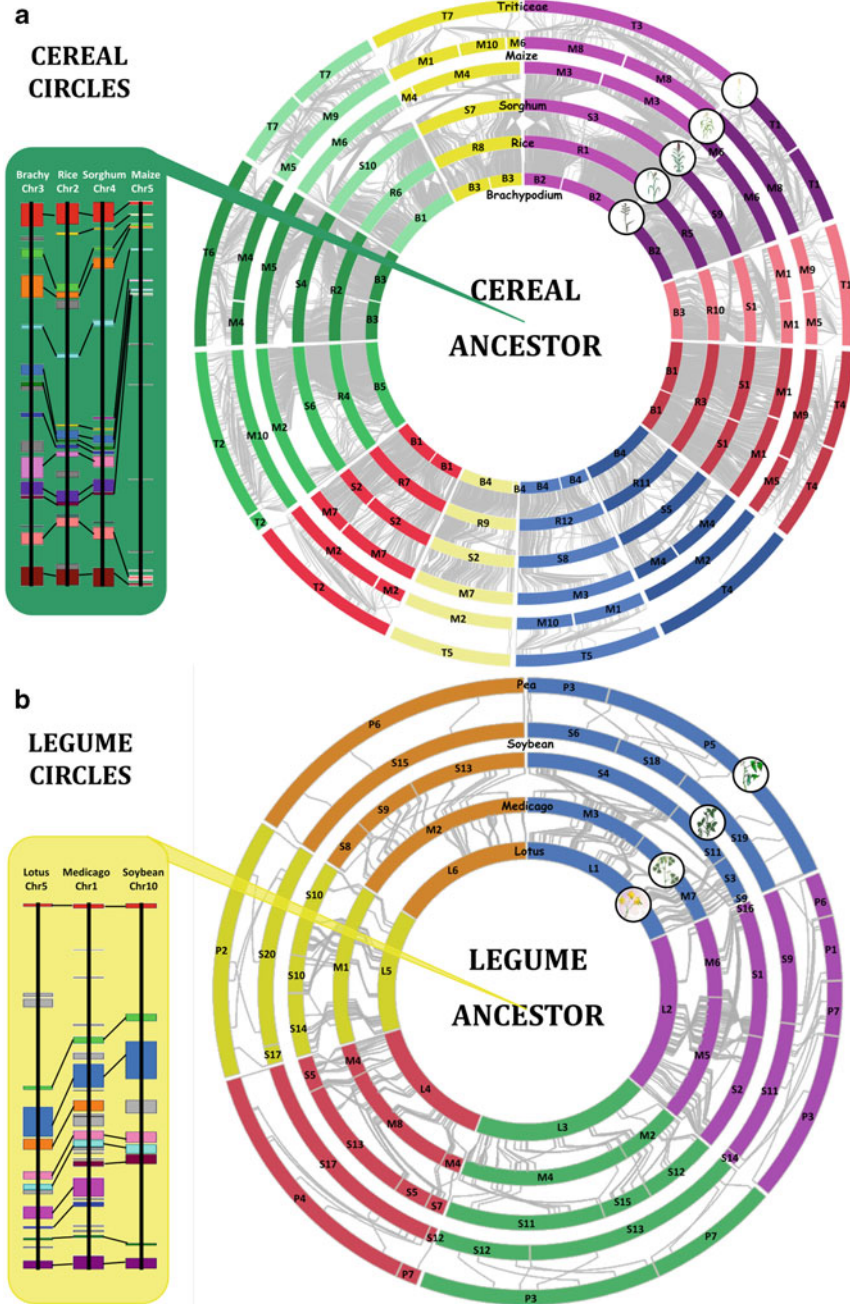


Fig. 7.4 Crop Circles (Adapted from Salse and Feuillet 2011)—The chromosome-to-chromosome synteny are represented as concentric circles for the cereals (rice, Brachypodium, sorghum, maize, Triticeae in panel **a**) and legumes (Lotus, Medicago, soybean, pea in panel **b**). The color code reflects the ancestral karyotype of the cereals ($n = 5$ ancestor) and legumes ($n = 6$ ancestor). The species associated with double circles (i.e. maize for the cereals and soybean for the

that these diploid plant species are all modern diploidized paleopolyploids (*cf* Table 7.3 and Paterson et al. 2004; Salse et al. 2008a; Salse 2012). Despite ancestral shared polyploidization events, additional species- or lineage-specific WGDs has been reported as summarized in Table 7.3. Figure 7.2c illustrates schematically the characterisation of shared *vs* specific whole genome duplication events. Ancestral duplications correspond to paralogous blocks within genomes that are located at orthologous positions between genomes. In contrast, lineage-specific duplications correspond to duplicated blocks within a genome that are orthologous to a single unique block in other genomes.

7.1.3.3 Ancestral Plant Karyotypes

The ancestral karyotype also called pan-genome consists in a core genome made of genomic features (chromosomes and genes) that are common to all investigated species (Salse et al. 2009b; Gavranović et al. 2011; Salse et al. 2011; Salse 2012). Conversely, the dispensable genome consists in the genomic features that are specific to each species such as recent gene transposition and/or TE bursts. The characterisation of seven shared paleoduplications and the precise relationships between different conserved regions allowed us to identify unlinked Conserved Ancestral Regions (CARs) or protochromosomes as well as evolutionary events that have shaped the modern plant genomes since their divergence from the founder karyotypes. Regarding the monocots, 50–90 million years ago, the $n = 5$ ancestor (AGK for Ancestral Grass Karyotype) went through a WGD (1R ancestral), followed by two chromosome fissions (hereafter *Cfis*) and fusions (hereafter *Cfus*) that resulted in an $n = 12$ ancestral intermediate ($5 + 5 + 2 = 12$ protochromosomes), Fig. 7.5 An alternative scenario illustrated on the Fig. 7.5 has been proposed with a $n = 7$ ancestor (Abrouk et al. 2010; Salse 2012). It is then possible to write the modern species chromosome number as an adequation based on 12 chromosomes (ancestral monocot karyotype intermediate) followed by $-X$ (X number of *Cfus*), $+Y$ (Y number of *Cfis*), and $\times 2^z$ (Z number of WGD). The modern monocots have derived from this $n = 12$ intermediate and consequently evolved through numerous rounds of ancestral chromosome fusion (*Cfus*), fissions (*Cfis*) and species-specific WGD (for maize) so that according to the previous mathematical rule: rice (12 chromosomes = 12), maize (10 chromosomes = $[12 - 2] \times 2 - 10$), *Brachypodium* (5 chromosomes = $12 - 7$), sorghum (10 chromosomes = $12 - 2$) and Triticeae (7 chromosomes = $12 - 5$), (Murat et al. 2010; Salse 2012).

Fig. 7.4 legumes) evolved through lineage specific WGD. Microsynteny illustration are provided at the left end side of the figure for the cereal [between rice (chromosome 2, 144 Kb, 18 genes), sorghum (chromosome 4, 123 Kb, 16 genes), maize (chromosome 5, 1 Mb, 18 genes), *Brachypodium* (chromosome 3, 119 Kb, 16 genes)] and legume [between Lotus (chromosome 5, 143 Kb, 15 genes), Medicago (chromosome 1, 138 Kb, 25 genes), soybean (chromosome 10, 103 Kb, 9 genes) circles. Conserved genes are linked with black lines and are of the same colours.

Similarly, the identification of at least remnants of the hexaploidy events (i.e. genome triplications) in all the eudicot genomes analysed favours the model with a hexaploid ancestor (1R ancestral), with an $n = 21$ ancestor intermediate common to all eudicots (Fig. 7.5, Abrouk et al. 2010; Proost et al. 2011; Salse 2012). In this scenario and according to the previous mathematical rules (based on WGD, *Cfus*, *Cfis* event) the grape, fragaria, cacao, lotus, soybean, apple, poplar, *Arabidopsis* and papaya can be written respectively as 19 chromosomes ($= 21 + 2 - 4$), 7 chromosomes ($= 21 + 9 - 23$), 10 chromosomes ($= 21 + 2 - 13$), 6 chromosomes ($= [21 + 1 - 16] \times 2 - 6$), 20 chromosomes ($= [21 + 1 - 16] \times 2 \times 2 + 13 - 17$), 17 chromosomes ($= [21 + 3 - 15] \times 2 + 4 - 5$), 19 chromosomes ($= [21 + 6 - 15] \times 2 + 4 - 9$), 5 chromosomes ($= [21 + 10 - 22] \times 2 - 13$), 9 chromosomes ($= 21 + 6 - 18$), cf Fig. 7.5.

Collectively, recent paleogenomic analyses, based on the accurate comparative genomics inferences of plant genomes described previously, allowed for the construction of an ancestral karyotype or pan-genome (<http://www.clermont.inra.fr/umr1095/plant-ancestor>) with a minimal physical size of $\sim 30\text{--}50$ Mb, structured in five/seven (monocots) or seven (eudicots) protochromosomes and comprising a minimum of $\sim 10000\text{--}15000$ protogenes (also referenced as ohnologs), Abrouk et al. 2010; Murat et al. 2010; Murat et al. 2012; Salse 2012. We consider that paleogenomics data can be considered as a guide for translational biology in plants. Considering an ancestral locus, up to 3 paralogous genes are observed in grape, 6 in poplar and 12 in *Arabidopsis*. In the same way, a unique ancestral gene for the monocot may correspond to two gene copies in rice, *Brachypodium*, sorghum, 4 in maize and 6 in bread wheat. However, in most cases, as discussed in the next Section (6.3.5), duplicated copies may have been lost or modified in their structure/function during evolution so that few ones may have retained the ancestral function. Consequently, an efficient translational genomic approach in transferring markers or candidate genes (putatively ‘functional ohnologs’) from one species (i.e. model species) to another (i.e. agronomic species), using paleogenomics or comparative genomics data, needs to consider all orthologous and paralogous regions between genomes of interest to unravel real ‘functional orthologs’.

7.1.3.4 Paleohistorical Shuffling Events

Despite the precise characterisation of ancestral plant karyotypes, precise evolutionary shuffling events have been identified including polyploidy (WGD), SD (segmental duplication), translocation, fusion, fission and deletion. Based on the precise datation of ancestral shared or recent specific genome-wide duplication events it has been proposed that paleopolyploidy events, usually considered as a rare and evolutionary dead-end phenomenon, may have been the basis for species diversification and even survival during the mass species extinction periods (Van de Peer et al. 2009b). Dating of such duplication events clearly identifies four distinct rounds of WGD during plant evolution: γ (ancestral WGD, > 65 mya), α/β (family-specific WGD, $30\text{--}65$ mya), ρ (lineage/species-specific WGD, < 30 mya), cf Figs. 7.1 and 7.5 (red dots). The ancestral paleoploidization event in monocot and eudicots can be

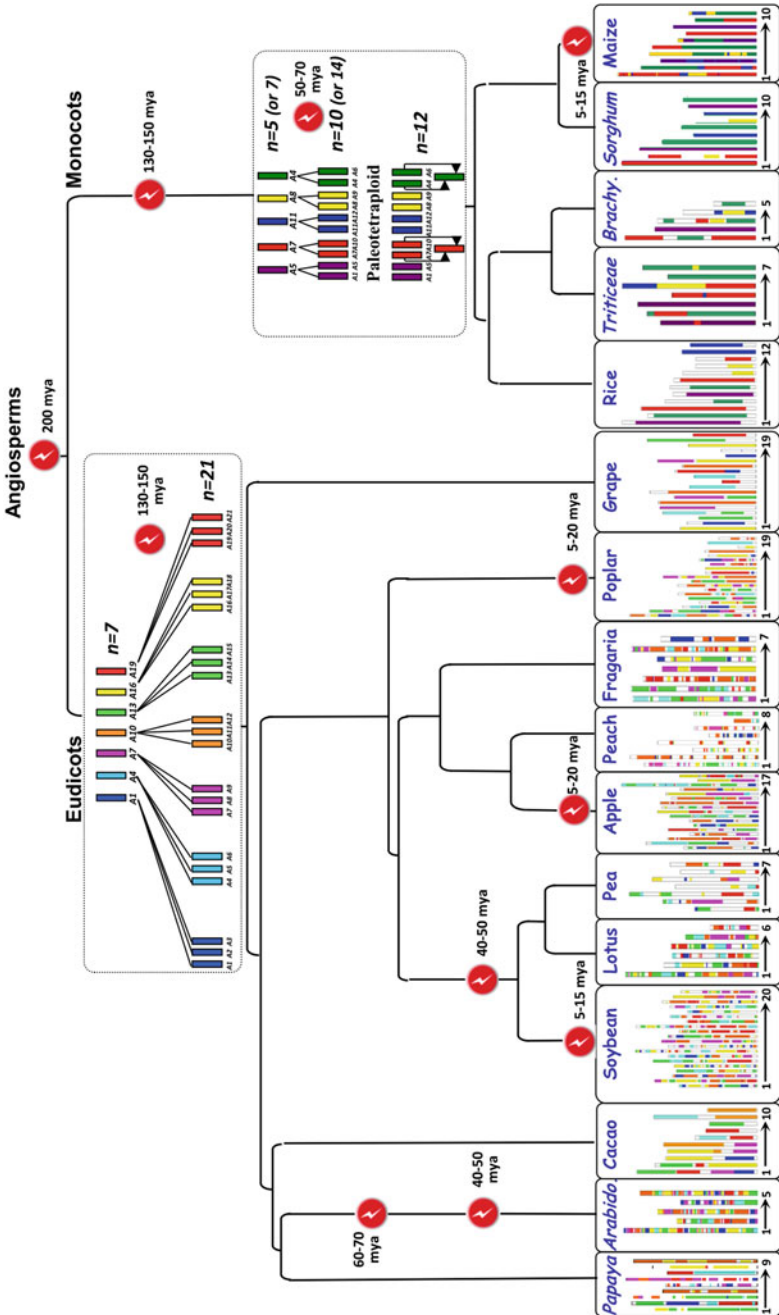


Fig. 7.5 Plant genome paleohistorical scenario (Adapted from Salse 2012). The present day monocot (*right*) and eudicot (*left*) genomes (*bottom*) are represented with color codes to illustrate the evolution of segments from their founder ancestors (*top*) with seven (*eudicots*) and five-seven

respectively associated with the Cretaceous/Paleogene (called K-PgT, 65 Mya) and also the Triassic/Jurassic (called Tr-J, 200 Mya) extinction events, Fig. 7.1 (Fawcett et al. 2009; Murat et al. 2012). Finally, all the reported recent and lineage specific WGD events are precisely associated with the Paleogene and Neogene periods. This geological period has been subjected to a drastic global cooling that occurred approximately 33 mya when tropical/sub-tropical adapted species from warm or humid climates have been reported to evolve rapidly to a more arid and cool condition (Morley 2003). Finally, it has been speculated that polyploidization in plants may be considered as an evolutionary process to face global climate or more generally environmental changes by providing putative selective advantages over their diploid (pre-duplication) progenitors (Comai 2005; Rieseberg and Willis 2007; Fawcett et al. 2009; Salse 2012). Overall, among the plant genome sequence available at the moment, grape and rice can be considered as the referenced genomes that still, in their modern structure, resemble the most to the ancestral founder karyotypes, respectively for the eudicots and monocots.

7.1.3.5 Structural and Functional Consequences of Evolution

More than 30 years ago, based on few protein sequences from vertebrates, Susumu Ohno proposed polyploidization as a major source of *de novo* biological pathways inherited from duplicated gene copies. Recent comparative genomics analyses in plants confirmed and refined Ohno's conclusion (Ohno 1970) and led to the identification of a polyploid common ancestor showing that the modern species have been shaped through several rounds of WGDs (Tang et al. 2008a, 2008b, 2010; Van de Peer et al. 2009a). Such polyploidizations generated a vast functional gene redundancy. This redundancy is rapidly reduced through the structural and/or functional divergence of the duplicates as a source of evolutionary novelty, suggesting that WGDs may have favoured species adaptation *via* structural/functional modulation of duplicated genes (Sankoff et al. 2010).

Polyploidization events are then followed by a diploidization process that acts at both genome-wide (duplicated chromosome fusion) and gene (duplicated gene shuffling) levels. At the genome wide level, polyploidizations followed by diploidizations provided a new dynamic pathway for extensive chromosome reshuffling based on chromosome fusions resulting in reduced numbers of chromosomes in today's plant species compared to their common paleopolyploid ancestors. Two types of ancestral chromosome fusion have been characterised and referenced as CCF (Centromeric Chromosome Fusion) and TCF (Telomeric Chromosome Fusion), *cf* Fig. 7.6a (Salse 2012). Such ancestral chromosome fusion events have led to dicentromeric chromosomes where one of the two centromeres is entirely deleted (in the case of the CCF) or inactivated (in the case of the TCF), Murat et al. 2010, Schubert and Lysak

Fig. 7.5 (*monocots*) protochromosomes (referenced as Ax). The WGD events that have shaped the structure of the different plant genomes during their evolution from the common ancestor are indicated as red dots

2011). When intervals comprising chromosome fusion points were compared in their structure, they may appear to correspond preferentially to (i) meiotic recombination hotspots, (ii) high sequence turn over loci through repeat invasion and (iii) hotspots of evolutionary novelty that could act as a reservoir for producing adaptive phenotypes (Murat et al. 2010). Finally, the modern genomes harbour in their actual chromosomal architecture traces of their evolutionary history and specifically of their specific patterns of ancestral chromosome fusions. A better understanding of the evolutionary processes (duplications and fusions) of plant chromosomes may allow us to develop in the next future appropriate tools aiming at monitoring, or maybe improving the evolution of modern species.

Paleopolyploid gene diploidization, in reducing duplicated gene redundancy, is also performed at the gene level, at both structural and expression/functional levels (Paterson et al. 2006; Coate and Doyle 2011). In term of structural shuffling, duplicated genes are lost by massive local deletion. However, duplicated gene deletion is not performed at random as it has been shown that duplicate gene redundancy is eliminated through the so-called sub-genome dominance phenomena (Paterson et al. 2010; Schnable et al. 2011) in which one of the duplicated blocks retain the majority of ancestral copy, whereas their duplicates are largely shuffled (deleted, transposed) in the sister block (Thomas et al. 2006; Woodhouse et al. 2010; Schnable et al. 2012). Recent evolutionary studies provide an opportunity to gain insight into the battery of genes that operated during the construction of modern plant species, especially those that have been structurally retained during evolution, referred and 'diploidization resistant' genes (Paterson et al. 2010; Freeling 2009). 'Diploidization resistant' gene families correspond to transcriptional regulators that are retained more significantly after WGD events and for which paralogous copies are maintained, leading to copy number variation (CNV, sequences that are present in different copy numbers between genomes). They contrast to 'diploidization sensitive' genes for which one paralogous copy is systematically lost to return to a diploid state leading to presence/absence variation (PAVs, sequences that are present in one genome, but absent in the other). Thus, additional CNVs of 'diploidization resistant' and PAVs of 'diploidization sensitive' genes with altered/modified functions would continually appear and be selected for during evolution, leading to colinearity erosion (Springer et al. 2009). The classes of retained genes after WGD (i.e. diploidization resistant) are often involved in dose-sensitive protein-protein interaction as member of multi-subunit complexes. Consequently, WGD may affect the kinetics and the function of the whole complex according to the 'Gene Dosage' hypothesis (Schnable et al. 2011; Freeling et al. 2009). Under this hypothesis, multi-partners complexes (such as FTs and TRs) are retained as duplicates where single-connected gene products are maintain as singletons, both phenomenon leading to the maintenance of the dosage balance following a WGD.

Duplicate genes that persist in multiple copies may diverge by differentiation of sequence and/or function. This process is affected by factors including pathway redundancy and modularity, as well as dosage of gene expression (Birchler and Veitia 2010; Bekaert et al. 2011). Overall, recurrent gene or genome duplications generate functional redundancy followed either by pseudogenization, concerted evolution,

subfunctionalization or neofunctionalization during the course of genome evolution (Fig. 7.6b). The derived functional divergence either from subfunctionalization or neofunctionalization processes between duplicated genes has been proposed as one of the most important sources of evolutionary innovation and plasticity in grasses (Throude et al. 2009; Pont et al. 2011). Finally, the consequence of polyploidization (reciprocal gene loss of paralogous gene copies or acquisition of novel functions. . .) could explain how WGD may have favoured the emergence of new gene functions and pathway and finally new plant species (Fig. 7.6b).

Several recent studies have precisely reported the impact of plant evolutionary events in general and genome duplications in particular on the agronomic traits elaboration. Table 7.5 provides a list of relevant studies establishing the role of WGDs, SDs and CNVs on major traits (i.e. flowering response, etiolation, pigmentation, photosynthesis, quality traits, yield, incompatibility) for several monocot and dicot species (sunflower, cotton, barley, Brassica, soybean, rice, wheat, Arabidopsis, maize).

One of the most relevant examples belongs to yield determinism in cereals where two maize GW (Grain Weight) genes belonging to chromosomal duplicates have been shown to control distinct phenotypic variations of kernel size and weight (Li et al. 2010). Despite yield components, it has also been suggested that genes retained after WGD are preferentially involved in response to biotic and abiotic changes. This has been clearly established through the subfunctionalization of two ancestrally duplicated CDPK (calcium dependant protein kinase) genes in wheat and rice, one evolving to ABA-specific response and the second to drought and cold responses (Geng et al. 2011). The other studies suggesting the evolutionary elaboration of agronomic traits through large or local duplication events are listed in Table 7.5. Finally, resistances to diseases is not reported in this table as the resistance genes are known to have evolved to recurrent as well as recent (i.e. lineage- or even species-specific) CNVs as will be reported in detail in the next Section (6.5.2), Leister et al. 1998.

7.1.4 CAR and Derived COS for Genetic and Physical Mapping

Comparative genomics and/or paleogenomics data may now provide a new reference such as ancestral karyotypes and derived Conserved Ancestral Regions (CARs) as an efficient platform for the development of universal high-throughput gene-based markers (COS for Conserved Orthologous Set) and accelerated translational genomics-based trait dissection in plants.

7.1.4.1 Computed Gene Order in Complex Non-Sequenced Genomes

Based on the precise identification of orthologous regions/segments between plant genomes and the derived CARs (see Section 6.3.3), it is then possible to perform a

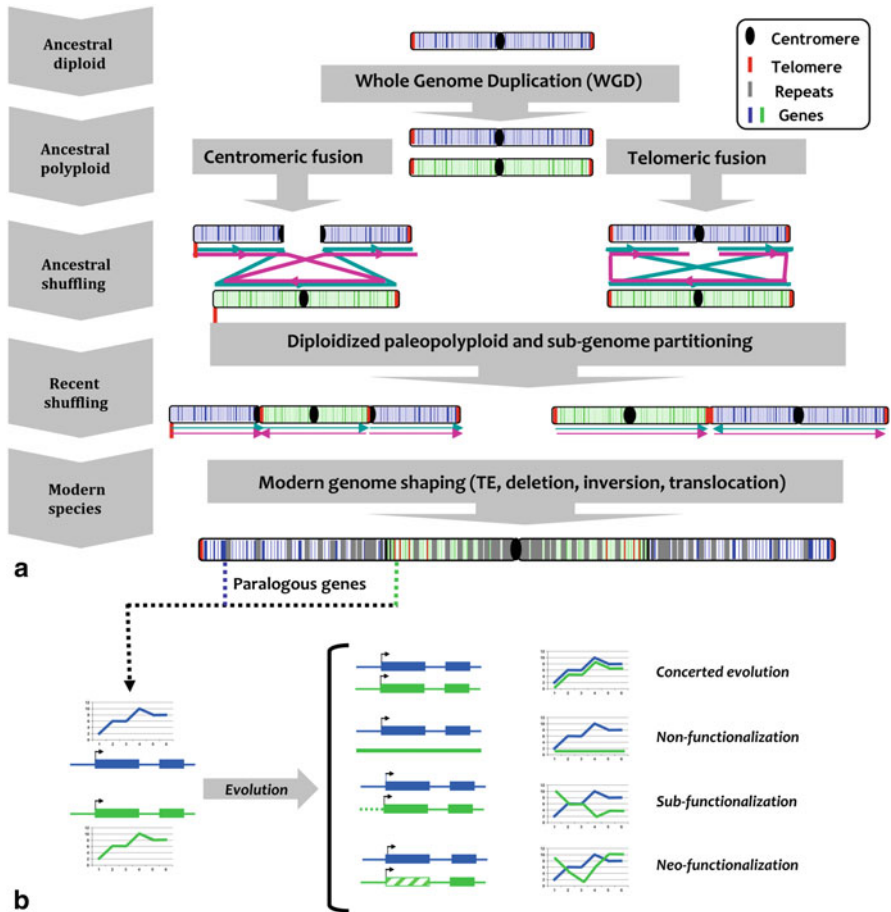


Fig. 7.6 Structural and functional consequences of WGD (adapted from Murat et al. 2010)—At the genome-wide level (panel **a**), the model begins with a single chromosome (*blue*) that is duplicated into a paralogous chromosome (*green*). The polyploidy state induces chromosome fusion either centromeric (*left*) or telomeric (*right*). Coloured arrows represent the different alternative orientations of the fusion events. The diploidized genome then evolved through numerous organizational shuffling events including, inversion, deletions, translocations, as well as TE invasion to reach the modern genome structure. At the gene level (panel **b**) any gene in any modern genome has been duplicated at least once during evolution. The ancestral sister copies have evolved either through concerted evolution (the ancestral expression pattern and/or function is retained in both copies), non-functionalization (with the complete deletion of one copy), sub-functionalization (with the modification of the original expression pattern and/or function) and finally neo-functionalization (with the acquisition of a new expression pattern and/or function for one of the copies)

multi-dimensional synteny-based approach to produce the most parsimonious computed (also referenced as simulated or virtual, Mayer et al. 2009) gene order in non-sequenced genomes based on gene conservation observed among the available

Table 7.5 Impact of genome duplication of trait elaboration—The table details the principal traits and associated loci (*or genes*) directly shaped by polyploidy (*WGD*), Segmental Duplication (*SD*) or Copy Number Variation (*CNV*)

Locus	Function	Plant Species	(WGD)-(SD)- (CNV)	References
Terminal flower	Flowering time	Sunflower	WGD	Blackman et al. 2011
CONSTANS	Flowering time	Barley	WGD	Cockram et al. 2010
NA	Fiber	Cotton	CNV	Rong et al. 2010 Zhu et al. 2011
FLC	Flowering time	Arabidopsis	SD	Nah et al. 2010 Rosloski et al. 2010
FLT	Flowering time	Brassica	SD	Wang et al. 2009
Rxp	–	Soybean	WGD	Kim et al. 2009
Ha	Grain quality	Wheat	CNV	Chantret et al. 2005
S locus	Self incompatibility	Brassica	CNV	Zhang et al. 2011
Oak	Incompatibility	Arabidopsis	CNV	Smith et al. 2011
–	Incompatibility	Arabidopsis	WGD	Bikard et al. 2009
–	Etiolation	Rice	WGD	Mao et al. 2011
GW2	Kernel size and weight	Maize	WGD	Li et al. 2010
SUN	Fruit shape	Tomato	SD	Xiao et al. 2008
P	Pigmentation	Maize	CNV	Chopra et al. 1998
C3/C4	Photosynthesis	Sorghum— Maize	WGD—CNV	Wang et al. 2009
CDPK	Stress response	Wheat—Rice	WGD	Geng et al. 2011

sequenced genomes (Murat et al. 2011; Pont et al. 2011). As an example, based on the chromosome-to-chromosome synteny relationships established between the seven bread wheat chromosome groups and rice, sorghum, *Brachypodium* and maize genomes, it is then possible to produce a partial wheat gene-based physical map where wheat sequences (EST, NGS data) were ordered within bread wheat chromosomes in respect to the position of their orthologous counterparts in rice, sorghum, *Brachypodium* and maize (Pont et al. 2011). Figure 7.7a illustrated the synteny-based computed gene order established upon the wheat chromosome 3B containing putatively 8400 genes (Choulet et al. 2010), where 3198 (38%) genes are available from the synteny-based and computed gene order strategy, with a density of more than one available synteny-based gene among three expected ones (Pont et al. 2011). Based on the percentage of conserved orthologous gene observed in plant genomes and described previously (see Section 6.3.1), we can assume that up to 80% of the computed gene order, in wheat for example, is exact and that lineage specific rearrangements, that cannot be extrapolated and then taken into account in such computed gene order approach, are limited and local.

However, this information can greatly increase the success of marker development (such as COS in the next section), because the selection of gene-based markers on the basis of the computed gene order is not based on a single reference genome

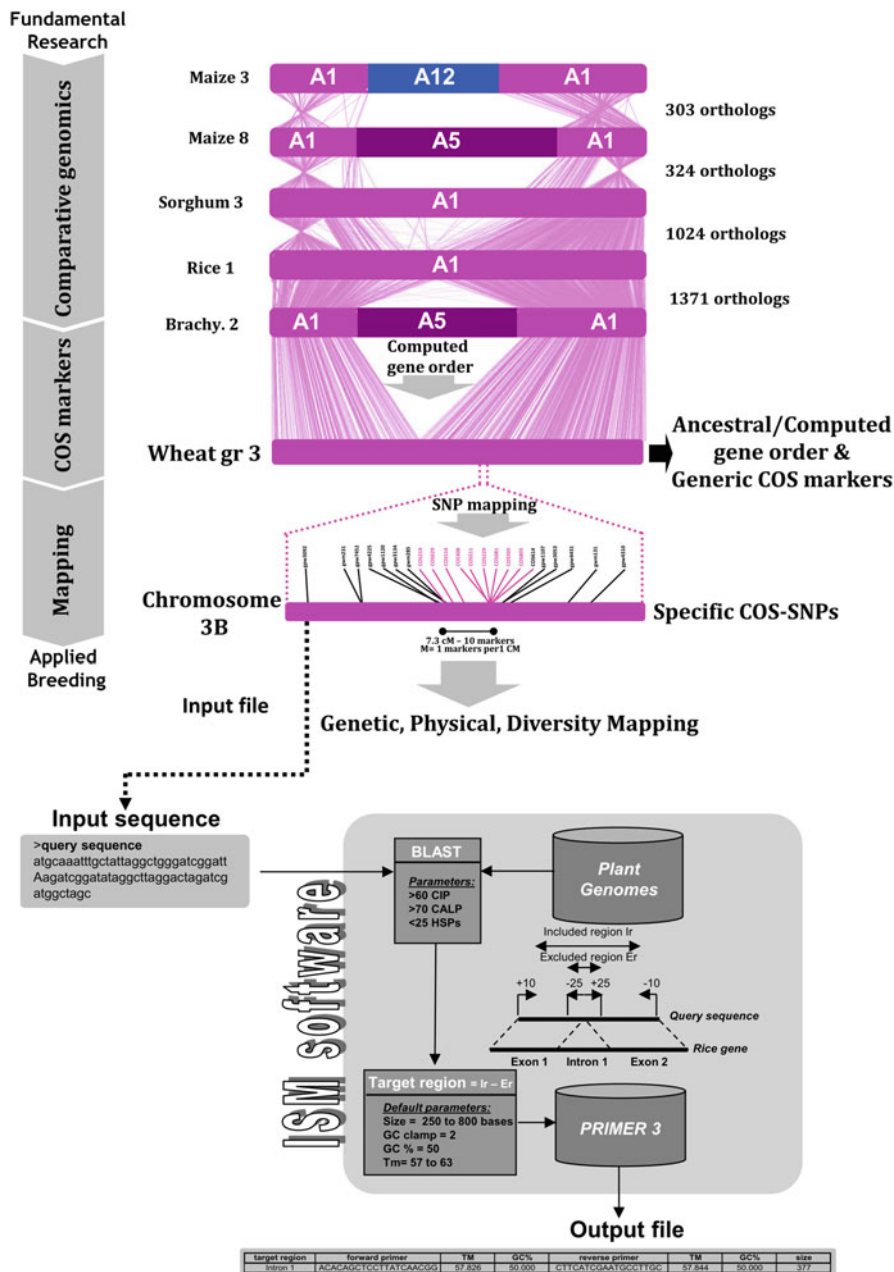


Fig. 7.7 Computed gene order and derived COS markers (adapted from Quraishi et al. 2009)—(a) Simulated synteny-based gene order model in bread wheat chromosome 3B (panel a) is illustrated based on the accurate synteny relationships (*top*) observed between maize (Chr 3-8), sorghum

(for example rice) and applied to another (wheat in this case example) with the risk that the locus of interest may have been subject to lineage-specific rearrangements in the unique model (rice) not shared with the target species (wheat). In contrast, Fig. 7.7a clearly illustrates that the delivered computed gene order derives from all the sequenced genomes available then not taking into account lineage-specific rearrangements but only the pan-genome conservation observed in the modern genome and then available in the reconstructed founder ancestors.

7.1.4.2 Universal Conserved Orthologous Set (COS) markers

Paleogenomics data provide information about the non-redundant ancestral plant gene set that can be used as a platform for the development of COS (Conserved Orthologous Set) markers (Fulton et al. 2002; Gupta and Rustgi 2004; Quraishi et al. 2009) to support cross genome (also referenced as syntenic) map-based cloning strategies. Paleogenomics data can greatly simplify and accelerate the identification of useful markers or candidate genes for a targeted chromosome locus (Fig. 7.7b). The relative organisation of the exons and introns is conserved across plant species, i.e. the number of exons and introns is maintained and individual introns occur at relatively the same sites for example between maize, wheat, *Brachypodium*, sorghum and barley orthologs. Exon conservation allows the development of intron spanning PCR-based primers located within conserved exons. A large set of COS markers suitable for plant genome mapping that are highly transferable (as derived from a robust synteny relationship between plants), highly polymorphic (as exploiting the largest source of polymorphism within introns, i.e. SNP), and co-dominant (as heterozygous haplotypes can be differentiated from homozygous ones) has been released in Quraishi et al. (2009) for the grasses, as well as associated tool for specific COS marker development: ISM (for Intron Spanning Marker) available at <http://www.clermont.inra.fr/umr1095/ISM>, for this purpose. These genic markers provide a rich and unlimited source of SNP polymorphism from which it is possible to create nearly ‘perfect’ markers or to saturate any genome or region of interest (Quraishi et al. 2011a, b).

Considering wheat as an example of COS marker development from any EST or gene-based NGS information (such as exome or RNA-seq for example), exon structures can be identified through wheat/rice-sorghum-maize-*Brachypodium* sequence alignments, as conserved HSPs correspond to exons. Precise exon/intron

Fig. 7.7 (Chr 3), rice (Chr 1) and *Brachypodium* (Chr 2) delivering an ancestral gene order in wheat (*bottom*). **(b)** Such conserved and ordered genes can be considered as a source of COS-SNP marker for genetic, physical as well as diversity mapping as illustrated at the bottom. The COS marker (primer and sequence) design software is schematically illustrated representing (i) the input sequence file at the top (nucleic sequence in fasta format) (i) the ISM (for Intron Spanning Markers) software principle at the center (alignment against rice genes with appropriate CIP and CALP values, selection of intron spanning primers with appropriate design criteria such as size, GC %, GC clamp, Tm), (iii) the output file at the bottom (table file with the primer name, sequence, Tm and GC %) values

Table 7.6 Plant markers databases—The table details the principal web services that deliver molecular markers for several plant species. Dedicated database to a unique species were not considered

Web	Markers	Plant species	References
PlantMarkers	SSR, SNP	50 species (EST-based)	http://markers.btk.fi/
GrainGenes	SSR, SNP	Triticeae—Avena	http://wheat.pw.usda.gov/GG2/index.shtml
Gramene	SSR, RFLP	Gramineae	http://www.gramene.org/
DArT	DArTs	> 50 species	http://www.diversityarrays.com/applicationsdart.html
AutoSNPdb	SNP	Barley—Brassica— Rice—Wheat	http://autosnpdb.qfab.org.au/
SOL	SSR, COS	Solanaceae	http://solgenomics.net/
GDR	SSR, SNP	Rosaceae	http://www.rosaceae.org/
BRAD	SSR	Brassicaceae	http://brassicadb.org/brad/
CGD	SNP, SSR	Cucurbitaceae	http://www.icugi.org/

boundaries (i.e. HSP boundaries) identified for any considered wheat sequence (including short reads from 454, Solexa, SOLID) associated with a rice-sorghum-maize-*Brachypodium* ortholog can be considered as templates to define two values, i.e. Ir and Er. Ir (for Included region) and Er (for Excluded region) are associated with any Intron position (I_i) within a wheat sequence aligned with a rice-sorghum-maize-*Brachypodium* sequence: Er [= ($I_i - 25$) to ($I_i + 25$)] corresponds to 50 nucleotides centred on the predicted intron position within the wheat EST sequence. Ir [= ($I_{i-1} + 10$) to ($I_{i+1} - 10$)] corresponds to the two consecutive exons spanning the predicted intron position within the wheat EST sequence. The precise sequence region corresponding to Ir-Er is provided to the Primer 3 package to select primer pairs on exons for intron (for SNP discovery purpose) or exon (for evolutionary purpose) amplification with the following parameters suitable for detection on Applied Biosystems (ABI) capillary sequencer: (i) Primer size (20 to 25 mer as default parameters), (ii) Amplicon size (between 250–800 bp as default parameters), (iii) Tm (between 57–63 as default parameters), (iv) GC clamp (equal to 2, i.e. a G or C at the 5' extremity as default parameters), (v) GC percentage (50 % as default parameters), Fig. 7.7b.

The universal COS markers designed to cover any plant genome of interest (upon the available ISM tool) extensively allow (i) identification of genes (or diagnostic markers) associated to complex traits based on QTL analysis (detailed in the next Section 6.4.3) (ii) exploitation of genetic diversity (in regard to the natural variation/diversity). Several databases provide information on molecular markers (SSR, SNP, COS, RFLP, DArT, *etc*) from multiple plant species as listed in Table 7.6.

7.1.4.3 Strategy of Physical and Genetic Mapping from Ancestral Karyotypes

Cloning genes driving traits of interest can be achieved either in sequenced and non-sequenced genomes. In both cases, paleogenomics or comparative genomics data are of major interest. Figure 7.8a illustrates the classical procedure followed to dissect and ultimately clone genes driving traits in sequenced genomes. The genome

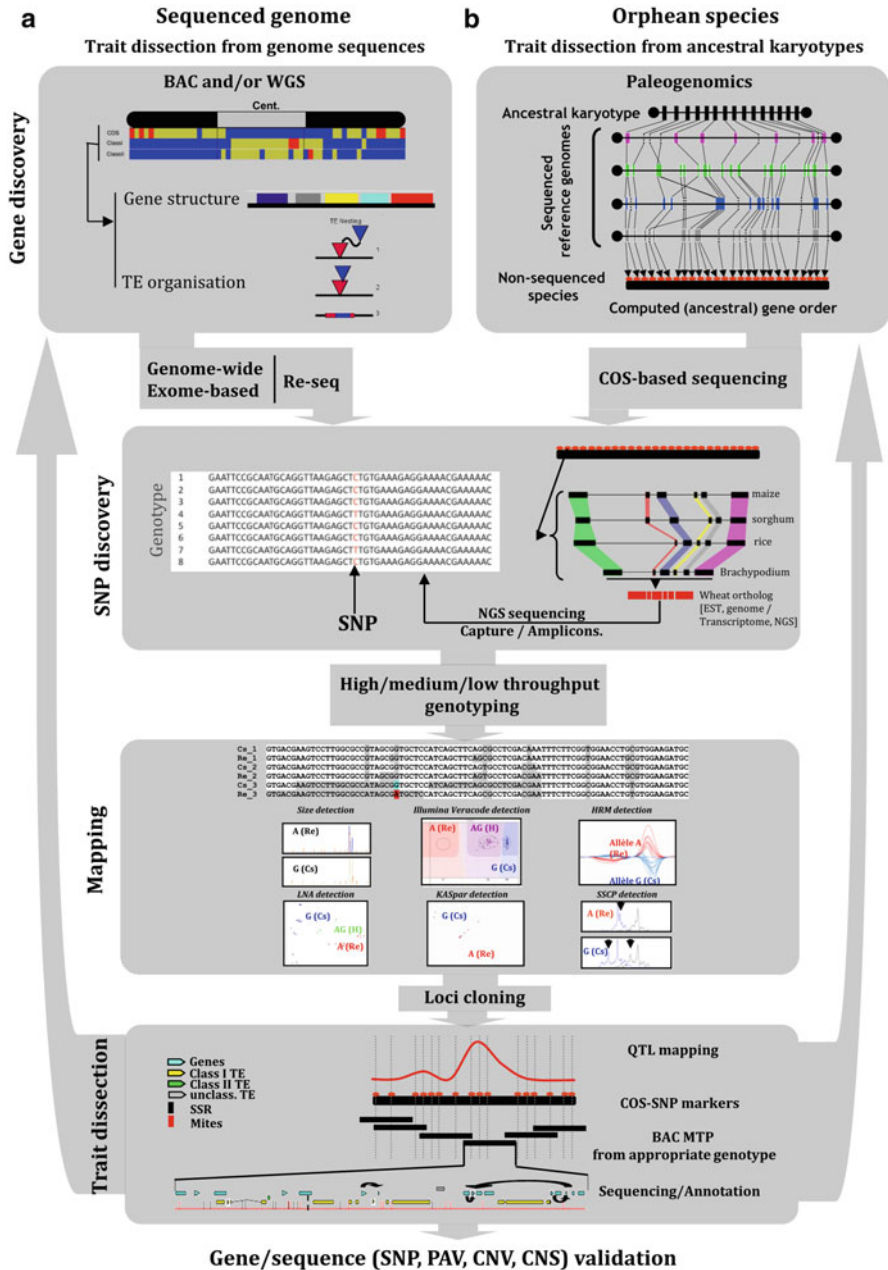


Fig. 7.8 Cross genome map based cloning strategy—Genome sequence-based trait dissection (panel a) and paleogenomics-based trait dissection (panel b) complementary strategies are illustrated according to gene discovery (from genome draft or paleogenomics information), SNP discovery (from sequence-based or COS markers), mapping (from Illumina, LNA, KASPar, HRM, SSCP technologies), trait dissection (from large scale intra-specific identification of causal SNPs, PAVs, CNVs) steps according to the text description

sequence deliver a non-limited set of molecular markers either genic (based on the gene model annotation) or intergenic (based on the transposable element annotation), see gene discovery panel. In this case, comparative genomics in delivering a robust set of orthologous regions in numerous species allow refining the gene models (intron/exon structures) for the species of interest when selecting genes for SNP development (see SNP discovery panel). Genome-wide and exome-based re-sequencing in a large panel of genotypes deliver a large set of SNP that can be mapped using classical genotyping methods such as Illumina, LNA, KASpar, HRM, SSCP (see mapping panel).

The reduction of the genetic interval of trait is obtained by integrating results obtained from different segregating populations (i.e. meta-QTL from RIL, NIL, HIF, AB-QTL material) and/or by using historical patterns of recombination (i.e. association mapping from diversity panel). The access to the genome sequence allows one to obtain, in theory, a marker on each candidate sequence of interest within the genetic interval harbouring the trait. However, the genome sequence available for the genetic interval of interest comes, most of the time, from a single genotype selected for the genome sequencing initiative that may often not harbour the traits (i.e. favourable alleles or even the driving gene or sequence) of interest. In the case of paleogenomics-driven trait dissection in sequenced or even non-sequenced genomes, the use of comparative genomics data is central. Figure 7.8b illustrates such complementary strategy where the sequencing initiative is not initial and genome-wide for a single genotype of interest (as previously reported) but in contrast local (locus-based) and genotype-specific (most favourable genotypes for the trait of interest). Paleogenomics data and the derived computed gene order within the species of interest deliver a tremendous source of COS markers for any CAR related to the trait loci (see gene discovery panel). Such COS markers can be sequenced (NGS approaches) either from amplification or capture approaches (see SNP discovery). Based on the comparative genomics data presented in the Section 6.3, up to 80 % of the genes can be then targeted for SNP discovery within non-sequenced genomes (as described in the mapping panel). Finally, the trait interval can be covered ultimately by one COS-SNP marker per gene for BAC anchoring. In this strategy, species-specific BAC libraries are produced for several genotypes associated with the trait (for example QTL or metaQTL). The sequencing effort takes place in this final step with a large-scale, intra-specific sequencing approach so that a complete set of putative causal SNPs, PAVs and CNVs, can be used as functional markers (see trait dissection panel).

High resolution and large-scale comparative genomics studies offer a tremendous set of gene-based markers that can be used directly as founder resource for genome mapping (physical or genetic) and ultimately trait dissection.

7.1.5 Complex Traits Dissection

Comparative genomics have been used during the last 20 years to transfer information and resources from models (generally rice for the monocots and Arabidopsis for the eudicots) to non-sequenced genomes for agronomic trait dissection, i.e. strategy

referenced as translational biology. We have now more than 30 relevant cases in grasses of such translational genomics approach regarding disease resistance, developmental and quality traits to draw some conclusions regarding the advantages and limits of such strategy.

7.1.5.1 Examples of Conserved and Non-Conserved Traits/Genes in Grasses

The relevance of translational genomics (i.e. use of colinearity between sequenced model genomes and agronomic relevant crops) can be assessed by the number of failures and successes upon its exploitation. In order to precisely illustrate the use of paleogenomics data in dissecting traits, two examples will be detailed here from the literature, i.e. case of synteny-based improvement for Nitrogen Use Efficiency (NUE locus, Fig. 7.9a) and grain Hardiness (*Ha* locus, Fig. 7.9b) in bread wheat.

Monitoring Nitrogen Use Efficiency (NUE) in plants is becoming essential to maintain yield while reducing fertilizer usage. Optimized NUE application in major crops is essential for long-term sustainability of agriculture production. The precise identification of 11 major chromosomal regions driving NUE has been reported recently in wheat for which key developmental genes such as *Ppd* (photoperiod sensitivity), *Vrn* (vernalisation requirement), and *Rht* (reduced height) are in co-location and can be considered as robust markers from a molecular breeding perspective (Quraishi et al. 2011b). Precise physical mapping, sequencing, annotation and candidate gene validation of a NUE metaQTL located on the wheat chromosome 3B allowed to propose that a glutamate synthase (*GoGAT*) gene that is conserved structurally and functionally at orthologous positions in the rice, sorghum and maize genomes may contribute to NUE in wheat and in other cereals, Fig. 7.9a (Quraishi et al. 2011). This example illustrates the direct use of comparative genomics or paleogenomics data for direct candidate gene identification in complex non-sequenced genomes.

Comparative sequencing was also performed at the *Ha* (Hardiness) locus that controls grain hardiness in wheat with contrasted conclusions. Orthologous BACs were compared in *Triticum aestivum*, *Triticum durum*, the diploid relatives *Triticum monococcum* and *Aegilops tauschii*, (Chantret et al. 2005) and rice-sorghum-maize-*Brachypodium* (Fig. 7.9b). Rearrangements, such as transposable element insertions, sequence deletions, duplications and inversions involving illegitimate recombination were shown to be responsible for the numerous shuffling events observed between the different grasses as well as observed in wheat at different ploidy levels. The results showed that *Pin* genes (*Pina*, *Pinb*), driving grain hardiness, (i) are absent in rice-sorghum-maize-*Brachypodium*, (ii) are conserved in wheat diploid progenitors (i.e. the D genome of *Ae. tauschii* and the A genome of *T. monococcum*), (iii) but are lacking in the A and B genomes of tetraploid wheat *Triticum turgidum*. These comparative genomics data provided the following evolutionary scenario in which the *Pina* and *Pinb* genes were removed through large deletions resulting from TE-mediated illegitimate recombination that have occurred independently in the A and B genomes following the tetraploidization 1.5 Mya making durum wheat

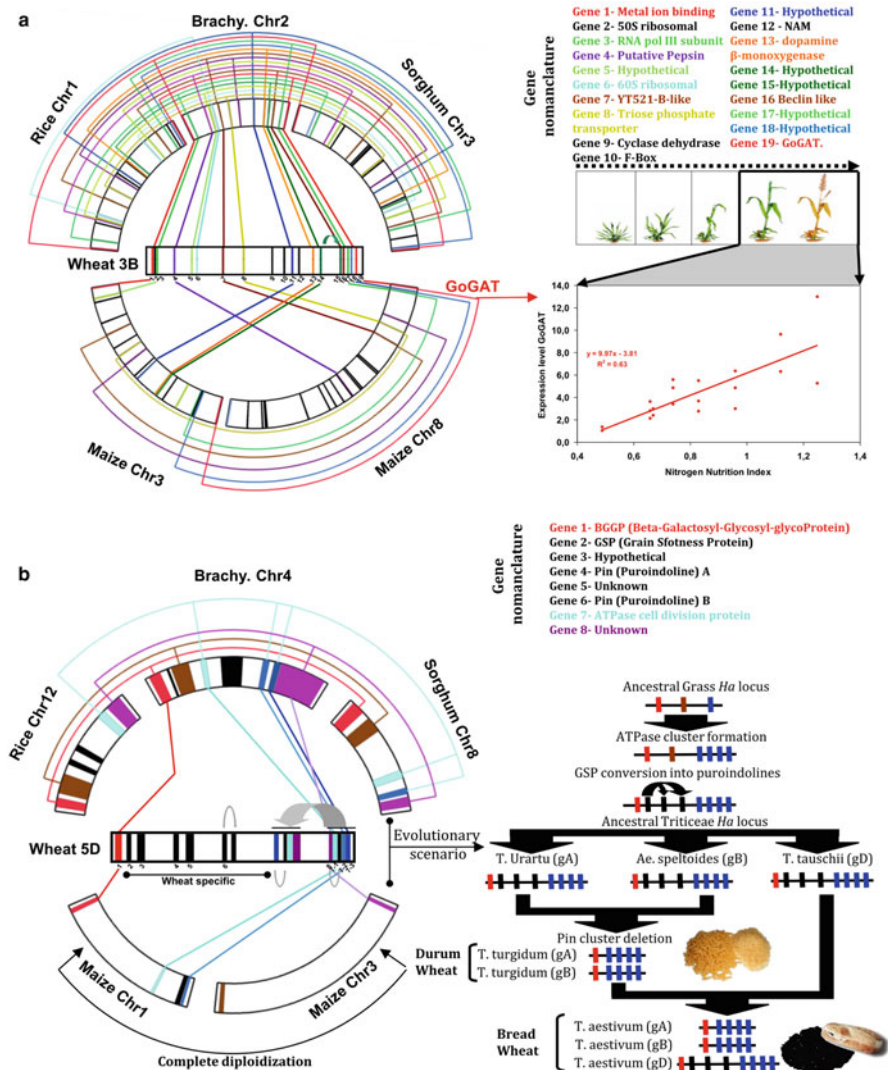


Fig. 7.9 Translational genomics in wheat (adapted from Quraishi et al. 2011b and Chantret et al. 2005)—The comparative sequence annotation at the NUE locus (panel a) between the 2.4 Mb sequence 3B region flanked by the two markers defining the NUE-QTL confidence interval and the rice, maize, sorghum and Brachypodium orthologous region is illustrated. Conserved genes (with the nomenclature at the top right) are linked with thin black lines. The thick red line illustrates the orthologous relationships observed for the GoGAT gene showing a linear regression observed between the GoGAT gene expression (expressed as $\Delta\Delta CT$ as y-axis) and the NNI status (as x-axis) for the Arche (red line) genotypes for 29 leaf samples (red dots) collected after flowering, bottom right. The microcolinearity studies at the Hardiness locus in cereal (panel b) is illustrated with conserved genes (nomenclature at the top right) linked with lines and with the same colour code between wheat (Chr 5D), maize (Chr 1-3), sorghum (Chr 8), rice (Chr 12) and Brachypodium

with hard grains, Fig. 7.9b. Finally, the second polyploidization event, 0.1 mya, between the tetraploid (*Pin* gene absent) and the D genome progenitor (*Ae. tauschii* harbouring *Pin* genes) delivered hexaploid bread wheat with soft grains (Chantret et al. 2005). This latter example illustrates a relevant case of trait specific dissection in bread wheat where comparative genomics established that the trait-causal gene has been acquired specifically in wheat lineages and has been subjected to intense organizational shuffling events during the last million year of evolution, then not conserved in the model genomes.

7.1.5.2 Comparative Genomics-Based Trait Dissection in Grasses

Translational genomics approach has been widely used in plants to identify causal for traits in non-sequenced genomes based on the orthologous regions identified in reference sequenced genomes. *Arabidopsis* and rice, the first plant genomes sequenced have been widely used as model genomes for such translational genomics strategy in eudicots and monocots, respectively. We will focus on translational genomics results obtained in grasses from which the two previous examples of conserved trait (NUE locus) and species-specific trait (Ha locus) have been described. One of the first microcolinearity studies was performed at the Shrunken 2/Anthocyaninless1 (*sh2/a1*) orthologous locus that was originally studied in maize, sorghum and rice (Chen et al. 1997, 1998). Despite large differences in the length of the intergenic regions in maize compared to rice and sorghum and a tandem duplication of one gene (*A1*) in sorghum, the linear order of the four genes (*Sh2*, *X1*, *X2* and *A1*) present at this locus was remarkably conserved between the three species. In contrast, in the Triticeae, colinearity was limited to the conservation of the *Sh2* and *X1* genes on chromosome 1L whereas the two other genes, *X2* and *A1*, were found on a non-orthologous chromosome (3L). This indicated that numerous rearrangements including gene translocations and transpositions have occurred at the locus since the divergence between the Triticeae and the other grasses (Li and Gill 2002). Since these first studies, several other micro-colinearity studies involving grass species have been performed at different loci carrying genes involved in disease resistance (e.g. *Lrk*, *Lr10-21-34*, *Pm3*, *vrs1*, *Rph7*, *mlo*, *Ror1-2*, *Rpg1-5*, *Rym4-5*, *Rar1*), plant development (e.g. *Vrn1-3*, *PhdH1*, *Ph1-2*, *Rht1*, *Q*, *Tga1*, *Vgt1*, *Ra1-3*, *BAF1/Bal*, *Th*, *Ts4-6*, *NUE*), and grain quality (e.g. *Ha*, *Glu*, *SPA*) traits as summarised in Table 7.7.

Generally, genes and associated QTLs involved in developmental processes (cf in Table 7.7 for Nitrogen assimilation, flowering time, architecture, photoperiodism, protein storage, etc) and that have been selected during evolution and/or domestication show a good level of conservation between grass genomes and the reconstructed

Fig. 7.9 (Chr 4). The proposed schematic evolutionary scenario of the Ha locus in cereals (bottom right) is illustrated with conserved flanking genes (BGGP in red and ATPases in blue), puroindolines (*Pin*) specific genes in black involving wheat diploid ancestors (*T. urartu*, *Ae. speltoides*, *T. tauschii*), tetraploid durum wheat (*T. durum*), hexaploid soft wheat (*T. aestivum*)

Table 7.7 Cross map-based trait dissection in cereals—The table details the cereal loci (*first column*) and traits (*second column*) dissected through translational genomics approach from mainly rice to non-sequenced species (*third column*) where driving gene was conserved or not (reference as C or nC in the *fourth column*)

Locus	Function	Plant Species	Conserved (C) Non-conserved (nC)	References
<i>NUE</i>	Nitrogen Uptake	Wheat	C	Quraishi et al. 2011
<i>Lrk</i>	Disease resistance	Wheat	NC	Feuillet and Keller 1999
<i>Pm3</i>	Disease resistance	Wheat	nC	Wicker et al. 2007 Yahiaoui et al. 2004
<i>Lr21</i>	Disease resistance	Wheat	nC	Brooks et al. 2002
<i>Lr10</i>	Disease resistance	Wheat	nC	Stein et al. 2000 Feuillet et al. 2003 Loutre et al. 2009
<i>Lr34</i>	Disease resistance	Wheat	nC	Bossolini et al. 2007
<i>mlo</i>	Disease resistance	Wheat	C	Elliott et al. 2002
<i>Rht1</i>	Dwarfing	Wheat	C	Ellis et al. 2002
<i>Vrn1</i>	Flowering	Wheat	C	Ramakrishna et al. 2002 Yan et al. 2003
<i>Vrn2</i>	Flowering	Wheat	C	Yan et al. 2004
<i>Vrn3</i>	Flowering	Wheat	C	Yan et al. 2006
<i>Ph1</i>	Pairing	Wheat	C	Griffiths et al. 2006
<i>Ph2</i>	Pairing	Wheat	C	Sutton et al. 2003
<i>Q</i>	Domestication	Wheat	nC	Faris et al. 2008
<i>sh2/a1</i>	Grain	Wheat	C	Shen et al. 1997–1998 Li et al. 2002 Bennetzen and Ma 2003
<i>Ha</i>	Hardiness	Wheat	nC	Chantret et al. 2005
<i>Glutenin</i>	Storage protein	Wheat	C	Wicker et al. 2003 Gu et al. 2006
<i>SPA</i>	Storage protein	Wheat	nC	Salse et al. 2008b
<i>Ppd-H1</i>	Photoperiodism	Barley	C	Dunford et al. 2002 Turner et al. 2005
<i>Ror1</i>	Disease resistance	Barley	C	Freialdenhoven et al. 1996
<i>Ror2</i>	Disease resistance	Barley	C	Huchelhoven et al. 2000
<i>Rpg1/5</i>	Disease resistance	Barley	C	Killian et al. 2005 Brueggeman et al. 2002 Drader and Kleinhofs 2010
<i>Rym4/5</i>	Disease resistance	Barley	nC	Tyrka et al. 2008
–	Disease resistance	Barley	nC	Chen et al. 2005
<i>Rar1</i>	Disease resistance	Barley	nC	Piffanelli et al. 1999
<i>Rph7</i>	Disease resistance	Barley	nC	Brunner et al. 2003 Scherrer et al. 2005
<i>vrsl</i>	Spike architecture	Barley	C	Pourkheirandish et al. 2007
<i>Tga1</i>	Kernel architecture	Maize	C	Chuck et al. 2007
<i>Vgt1</i>	Flowering	Maize	C	Salvi et al. 2007
<i>Ral3</i>	Kernel architecture	Maize	C	Satoh-Nagasawa et al. 2006
<i>BAF1/Ba1</i>	Kernel architecture	Maize	C	Gallavotti et al. 2011)
<i>Ts4/6</i>	Kernel architecture	Maize	C	Chuck et al. 2007

ancestral genome, deliver, in this case, computed gene orders as good matrix for direct gene isolation. Among the 20 referenced cases of developmental trait dissection illustrated in Table 7.7, 80% (16) show causal or candidate gene conservation in models (i.e. sequenced genomes), mainly rice in these studies or reconstructed ancestral pan genomes. In contrast, other types of genes do not show colinearity between the grass genomes. Indeed, there are very few examples of colinearity retained for disease resistance R genes between grass genomes and so far, map-based cloning of R genes in plant was not significantly benefiting from the sequence genome information or comparative genomics data. The non-syntenic location of R genes in the cereals was described 15 years ago through a comparative genetic analysis of resistance gene analogs in barley, rice and foxtail millet (Leister et al. 1998). In many cases, the attempts to use colinearity with rice for isolating R genes have revealed the limits of colinearity between sequenced genomes and the other non-sequenced cereal genomes. The first note of caution was provided with the barley stem rust resistance gene *Rpg1* map-based cloning project. Despite a certain degree of colinearity retained at the orthologous locus in rice (Killian et al. 1997), no orthologous gene was present in the rice genome and classical positional cloning of *Rpg1* was performed in barley (Brueggeman et al. 2002). In some cases, such as the leaf rust *Lr10* and the powdery mildew *Pm3* fungal disease R genes on wheat chromosome 1AS, the rice genome contains gene homologs to the wheat candidate genes but at non-orthologous positions, indicating massive genome rearrangements (Guyot et al. 2004). Both R-genes were also cloned using mixed strategies integrating positional cloning and synteny-based information/markers (Stein et al. 2000; Feuillet et al. 2003; Yahiaoui et al. 2004; Table 7.7). Overall, among the 12 referenced cases of disease resistance trait dissection in Table 7.7, 20% (3) show partial causal or candidate gene conservation in the model genomes (i.e. sequenced genome).

We illustrated in the previous paragraph contrasted successes of paleogenomics-based trait dissection. However, even if the causal gene is not present at its orthologous position in the model genome or reconstructed ancestral genome (essentially regarding R-genes), flanking genes are often conserved so that they can be used as a matrix to provide a source of markers (such as COS) to saturate the targeted region for high resolution physical/genetic mapping.

7.1.5.3 From Paleogenomics Data to Traits Improvement

Overall, translational genomics have been successfully used in the last 10 years based on a single reference genome in monocots (rice) and eudicots (*Arabidopsis*) either for direct gene identification (mainly for developmental traits) or as a source of molecular markers (mainly for non-conserved traits such as disease resistance). The access as a large set of genome sequences as well as robust comparative genomics data and derived tools, including the computed or ancestral gene order delivered from paleogenomics data, will largely increased the success of such approach. Based on this conclusion, we illustrate a general procedure from paleogenomics data to trait improvement in Fig. 7.10.

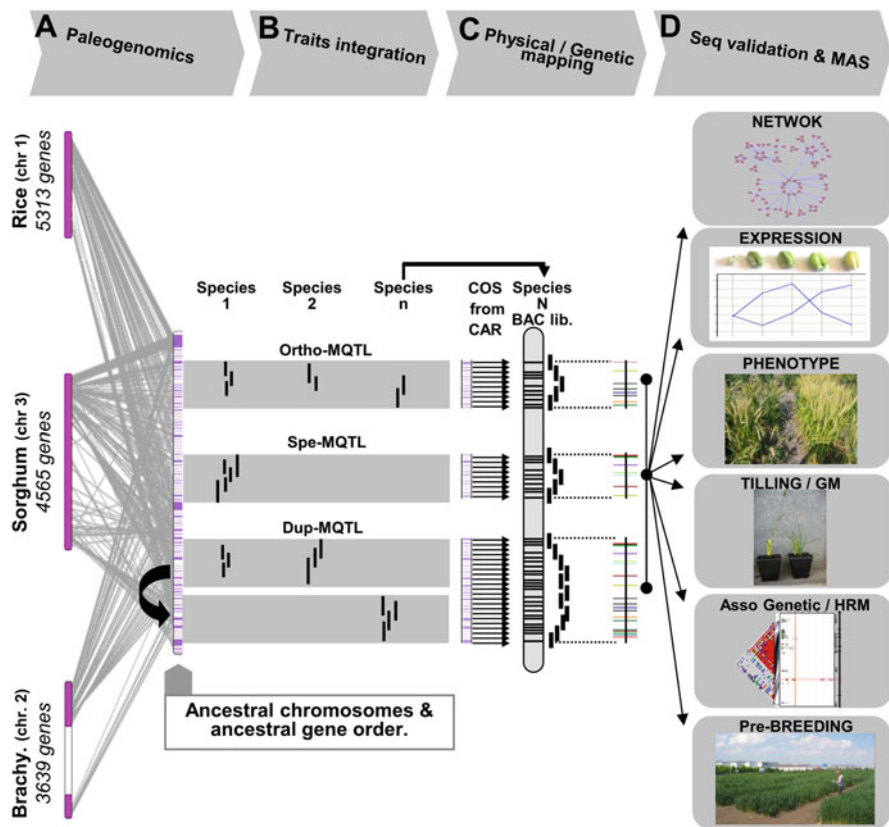


Fig. 7.10 From genome archaeology to crop improvement—The model schematically represents the steps in trait dissection following the paleogenomics (use of comparative genomics and computed ancestral gene order information), trait integration (identification of ortho-, spe- and para-MQTLs), mapping (physical and genetic, through NGS based sequencing), validation (network, expression, phenotype, tilling, HRM, pre-breeding informations) steps

The initial step consists in selecting traits and species that are favourable for translational genomics approaches by moving from comparative genomics or paleogenomics (panel a) to comparative genetics (panel b). It is now possible to compare quantitative genetics (i.e. association genetics and QTL/metaQTL) data in order to identify to an unprecedented resolution conserved traits (i.e. referenced as ortho-MQTL), species-specific traits (i.e. referenced as spe-MQTL) or alternatively, traits that have been shaped by chromosomal duplications (i.e. referenced as para-MQTL). Access to the paleogenomics basis of the evolution of key traits from modern cultivated crops referenced as ortho-MQTL, spe-MQTL or para-MQTL will benefit in selecting appropriate approaches in cloning causal genes or sequences. Based on the characterisation of the type/nature of traits investigated (ortho-MQTL or spe-MQTL or para-MQTL), the public paleogenomics studies deliver an exhaustive set of conserved genes as well as computed or ancestral gene order from which a complete and

exhaustive set of ordered COS markers can be selected genome-wide or locus-based. Such resources can be directly used as a direct candidate (mainly for developmental traits) identification or used for locus-based or targeted physical mapping construction (panel c). Sequence/gene validation is then classically achieved through the integration of information and data related to network, expression, phenotype, tilling, high resolution mapping, pre-breeding data (panel d).

Such translational or paleogenomics-based approaches need to be advanced to include, in the coming years, high resolution interactome/proteome/transcriptome/methylome data in order to transfer between species not only structural genomics data but functional ones. As a conclusion, the field of comparative genomics enters into a new area that we can consider as ‘functional paleogenomics’, where not only organizational relationships are made available between genomes/species but also functional homologs, that will be accurately identified for crop improvement in the so-called ‘translational biology’ inference.

7.2 Future Challenges

The past decade has seen a revolution in structural and functional genomics and has demonstrated the power of comparative studies in economically essential crop species. Comparative studies has led to improved genetic and physical maps, the development of large sets of accurate markers for breeding and the map-based isolation of a large set of genes of agronomic interest. It also provided insight into the evolution of plant, unravelling some of the major mechanisms that have shaped their genome within 150–300 million years of speciation. Comparative genomic studies between plant genome sequences will deliver additional information about plant genome evolution and better tools for crop improvement. There is no doubt that with the ongoing efforts, comparative studies will continue to provide invaluable information for a better understanding of the adaptation of plants to their environment and open new areas for breeding strategies, plant protection and conservation of biodiversity.

Acknowledgments The illustrations used in the chapter are derived from published studies supported by grants from the Agence Nationale de la Recherche, Program ANRjc entitled PaleoCereal (*ref*: ANR-09-JCJC-0058-01) and Program ANR-blanc entitled PAGE (*ref*: ANR-2011-BSV6-00801). The author would like to thank Caroline Pont and Florent Murat for their contribution in preparing the illustrations.

References

- Abrouk M, Murat F, Pont C et al (2010) Palaeogenomics of plants: synteny-based modelling of extinct ancestors. *Trends Plant Sci* 15(2010):479–487
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814):796–815
- Argout X, Salse J, Aury JM et al (2011) The genome of *Theobroma cacao*. *Nat Genet* 43(2):101–108

- Bekaert M, Edger PP, Pires JC et al (2011) Two-phase resolution of polyploidy in the arabidopsis metabolic network gives rise to relative and absolute dosage constraints. *Plant Cell* 23(5):1719–1728
- Bennetzen JL, Ma J (2003) The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. *Curr Opin Plant Biol* 6(2):128–133
- Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev* 15(6):621–627
- Bikard D, Patel D, Le Metté C et al (2009) Divergent evolution of duplicate genes leads to genetic incompatibilities within *A. thaliana*. *Science* 323(5914):623–626
- Birchler JA, Veitia RA (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186(1):54–62
- Blackman BK, Rasmussen DA, Strasburg JL et al (2011) Contributions of flowering time genes to sunflower domestication and improvement. *Genetics* 187(1):271–287
- Bolot S, Abrouk M, Masood-Quraishi U et al (2009) The ‘inner circle’ of the cereal genomes. *Curr Opin Plant Biol* 12(2):119–125
- Bordat A, Savoie V, Nicolas M et al (2011) Translational genomics in legumes allowed placing *in silico* 5460 Unigenes on the pea functional map and identified candidate genes in *Pisum sativum* L. *Genes, genomes, genetics*. doi:10.1534/g3.111.000349
- Bossolini E, Wicker T, Knobel PA et al (2007) Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J* 49(4):704–717
- Brooks SA, Huang L, Gill BS et al (2002) Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. *Genome* 45(5):963–972
- Brueggeman R, Rostoks N, Kudrna D et al (2002) The barley stem rust-resistance gene Rpg1 is a novel disease-resistance gene with homology to receptor kinases. *Proc Natl Acad Sci USA.* 99:9328–9333
- Brunner S, Keller B, Feuillet C (2003) A large rearrangement involving genes and low copy DNA interrupts the microcolinearity between rice and barley at the *Rph7* locus. *Genetics* 164:673–683
- Chantret N, Salse J, Sabot F et al (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* 17:1033–45
- Chen M, Bennetzen JL (1996) Sequence composition and organization in the Sh2/A1-homologous region of rice. *Plant Mol Biol* 32(6):999–1001
- Chen M, SanMiguel P, de Oliveira AC et al (1997) Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. *Proc Natl Acad Sci USA.* 94:3431–3435
- Chen M, SanMiguel P, Bennetzen JL (1998) Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. *Genetics* 148:435–443
- Chen H, Wang S, Xing Y et al (2005) Comparative analyses of genomic locations and race specificities of loci for quantitative resistance to *Pyricularia grisea* in rice and barley. *Proc Natl Acad Sci USA.* 100:2544–2549
- Chopra S, Athma P, Li XG et al (1998) A maize Myb homolog is encoded by a multicopy gene complex. *Mol Gen Genet* 260(4):372–380
- Choulet F, Wicker T, Rustenholz C et al (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22(6):1686–1701
- Chuck G, Meeley R, Irish E et al (2007a) The maize tasselseed4 microRNA controls sex determination and meristem cell fate by targeting Tasselseed6/indeterminate spikelet1. *Nat Genet* 39(12):1517–1521
- Chuck G, Cigan AM, Saetern K et al (2007b) The heterochronic maize mutant corngrass1 results from overexpression of a tandem microRNA. *Nat Genet* 39(4):544–549
- Coate JE, Doyle JJ (2011) Divergent evolutionary fates of major photosynthetic gene networks following gene and whole genome duplications. *Plant Signal Behav.* 6(4):594–597

- Cockram J, Howells RM, O'Sullivan DM (2010) Segmental chromosomal duplications harbouring group IV CONSTANS-like genes in cereals. *Genome* 53(3):231–240
- Comai L (2005) The advantages and disadvantages of being polyploid. *Nat Rev Genet* 6(11):836–846
- Courcelle E, Beausse Y, Letort S et al (2008) Narcisse: a mirror view of conserved syntenies. *Nucleic Acids Res* 36 (Database issue):D485–490
- Delseny M, Han B, Hsing YLe (2010) High throughput DNA sequencing: the new sequencing revolution. *Plant Sci* 179(5):407–422
- Devos KM, Gale MD (1997) Comparative genetics in the grasses. *Plant Mol. Biol* 35:3–15
- Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. *Plant Cell* 12:637–646
- Devos KM (2005) Updating the 'crop circle'. *Curr Opin Plant Biol* 8:155–162
- Drader T, Kleinhofs A (2010) A synteny map and disease resistance gene comparison between barley and the model monocot *Brachypodium distachyon*. *Genome* 53(5):406–417
- Dunford RP, Yano M, Kurata N et al (2002) Comparative mapping of the Barley *Ppd-H1* photoperiod response gene region, Which lies close to a Junction between two Rice linkage segments. *Genetics* 161:825–834
- Egan M, Lee EK, Chiu JC et al (2009) Gene orthology assessment with OrthologID. *Methods Mol Biol* 537:23–38
- Elliott C, Zhou F, Spielmeier W et al (2002) Functional conservation of wheat and rice *Mlo* orthologs in defense modulation to the powdery mildew fungus. *Mol Plant Microbe Interact.* 15(10):1069–1077
- Ellis H, Spielmeier W, Gale R et al (2002) "Perfect" markers for the Rht-B1b and Rht-D1b dwarfing genes in wheat. *Theor Appl Genet* 105(6–7):1038–1042
- Faris JD, Zhang Z, Fellers JP et al (2008) Micro-colinearity between rice, *Brachypodium*, and *Triticum monococcum* at the wheat domestication locus Q. *Funct Integr Genomics* 8(2):149–164
- Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc Natl Acad Sci U S A.* 106(14):5737–5742
- Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl Acad Sci USA.* 96:8265–8270
- Feuillet C, Travella S, Stein N et al (2003) Map-based isolation of the leaf rust disease resistance gene Lr10 from the hexaploid wheat (*Triticum aestivum* L.) genome. *Proc Natl Acad Sci U S A.* 100(25):15253–15258
- Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu Rev Plant Biol* 60:433–453
- Freialdenhoven A, Peterhansel C, Kurth J et al (1996) Identification of genes required for the function of non-race-specific mlo resistance to powdery mildew in barley. *Plant Cell* 8(1):5–14
- Fu H, Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci U S A.* 99(14):9573–9578
- Fulton TM, Van der Hoeven R, Eannetta NT et al (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14(7):1457–1467
- Gale M, Devos KM (1998) Comparative genetics in grasses. *Proc Natl Acad Sci U S A.* 95:1971–1974
- Gallavotti A, Malcomber S, Gaines C et al (2011) BARREN STALK FASTIGIATE1 is an AT-hook protein required for the formation of maize ears. *Plant Cell* 23(5):1756–1771
- Gavranović H, Chauve C, Salse J et al (2011) Mapping ancestral genomes with possible massive gene loss: a matrix sandwich problem. *Bioinformatics* 27(13):i257–265
- Geng S, Zhao Y, Tang L et al (2011) Molecular evolution two duplicate CDPK genes. *Gene* 475(2):94–103
- Griffiths S, Sharp R, Foote TN et al (2006) Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature.* 439:749–752

- Gu YQ, Salse J, Coleman-Derr D et al (2006) Types and rates of sequence evolution at the high-molecular-weight glutenin locus in hexaploid wheat and its ancestral genomes. *Genetics* 174(3):1493–1504
- Gupta PK, Rustgi S (2004) Molecular markers from the transcribed/expressed region of the genome in higher plants. *Funct Integr Genomics* 4(3):139–162
- Hampson S, McLysaght A, Gaut B et al (2003) LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res* 13:1–12
- Hückelhoven R, Trujillo M, Kogel KH (2000) Mutations in *Ror1* and *Ror2* genes cause modification of hydrogen peroxide accumulation in mlo-barley under attack from the powdery mildew fungus. *Mol Plant Pathol.*: 1(5):287–292
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- International *Brachypodium* Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463(7282):763–768
- Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467
- Jiao Y, Wickett NJ, Ayyampalayam S et al (2011) Ancestral polyploidy in seed plants and angiosperms. *Nature* 473(7345):97–100
- Kilian A, Chen J, Han F et al (1997) Towards map-based cloning of the barley stem rust resistance genes *rpg1* and *rpg4* using rice as an intergenomic cloning vehicle. *Plant Mol. Biol* 35:187–195
- Kim KD, Shin JH, Van K et al (2009) Dynamic rearrangements determine genome organization and useful traits in soybean. *Plant Physiol.* 151(3):1066–1076
- Kircher M, Kelso (2010) High-throughput DNA sequencing—concepts and limitations. *J Bioessays.* 32(6):524–536
- Leister D, Kurth J, Laurie DA et al (1998) Rapid reorganization of resistance gene homologues in cereal genomes. *Proc Natl Acad Sci USA.* 95:370–375
- Loutre C, Wicker T, Travella S et al (2009) Two different CC-NBS-LRR genes are required for Lr10-mediated leaf rust resistance in tetraploid and hexaploid wheat. *Plant J* 60(6):1043–1054
- Li W, Gill BS (2002) The colinearity of the Sh2/A1 orthologous region in rice sorghum and maize is interrupted and accompanied by genome expansion in the triticeae. *Genetics* 160:1153–1162
- Li Q, Li L, Yang X et al (2010) Relationship, evolutionary fate and function of two maize co-orthologs of rice GW2 associated with kernel size and weight. *BMC Plant Biol* 10:143
- Lyons E, Pedersen B, Kane J et al (2008) Finding and comparing syntenic regions among arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol.* 2008 Dec;148(4):1772–1781
- Mao D, Yu H, Liu T et al (2011) Two complementary recessive genes in duplicated segments control etiolation in rice. *Theor Appl Genet* 122(2):373–383
- Mayer KF, Taudien S, Martis M et al (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.* 151(2):496–505
- Ming R, Hou S, Feng Y et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature.* 452(7190):991–996
- Moore G, Devos KM, Wang Z et al (1995) Cereal genome evolution: grasses, line up and form a circle. *Current Biology* 5:737–739
- Morley JR (2003) Interplate dispersal paths for megathermal angiosperms. *Perspectives in Plant Ecology, Evolution and Systematics* 6(1–2):5–20
- Murat F, Xu JH, Tannier E et al (2010) Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res* 20:1545–1557
- Murat F, Peer YV, Salse J (2012) Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. *Genome Biol Evol.* 4(9):805–816
- Nah G, Jeffrey Chen Z (2010) Tandem duplication of the *FLC* locus and the origin of a new gene in arabidopsis related species and their functional implications in allopolyploids. *New Phytol.* 186(1):228–238

- Ostlund G, Schmitt T, Forslund K et al (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* 38(Database issue):D196–203
- Ohno S. (1970) Evolution by gene duplication. Springer-Verlag, Berlin, pp 160
- Pan X, Stein L, Brendel V (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics* 21(17):3461–3468
- Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci USA*. 101:9903–9908
- Paterson AH, Chapman BA, Kissinger JC et al (2006) Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis*, *Oryza*, *Saccharomyces* and *Tetraodon*. *Trends Genet* 22(11):597–602
- Paterson AH, Bowers JE, Bruggmann R et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457(7229):551–556
- Paterson AH, Freeling M, Tang H et al (2010) Insights from the comparison of plant genome sequences. *Annu Rev Plant Biol* 61:349–372
- Piegu B, Guyot R, Picault N et al (2006) Doubling genome size without polyploidization: dynamics of retrotransposon-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res* 16(10):1262–1269
- Piffanelli P, Devoto A, Schulze-Lefert P (1999) Defence signalling pathways in cereals. *Curr Opin Plant Biol* 2(4):295–300
- Pont C, Murat F, Confolent C et al (2011): Structural and functional consequences of neo- and paleopolyploidization events unveiled through RNA sequencing-based inference of grain gene network in bread wheat (*Triticum aestivum* L.). *Genome Biol* in press
- Pourkheirandish M, Wicker T, Stein N et al (2007) Analysis of the barley chromosome 2 region containing the six-rowed spike gene *vrs1* reveals a breakdown of rice-barley micro collinearity by a transposition. *Theor Appl Genet* 114(8):1357–1365
- Proost S, Van Bel M, Sterck L et al (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21(12):3718–3731
- Proost S, Pattyn P, Gerats T et al (2011) Journey through the past: 150 million years of plant genome evolution. *Plant J* 66(1):58–65
- Qi LL, Echalié B, Chao S et al (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 168:701–712
- Quraishi UM, Abrouk M, Bolot S et al (2009): Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. *Funct Integr Genomics* 9(4):473–484
- Quraishi UM, Murat F, Abrouk M et al (2011a) Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (*Triticum aestivum* L.). *Funct Integr Genomics* 11(1):71–83
- Quraishi UM, Abrouk M, Murat F et al (2011b) Cross-genome map based dissection of a nitrogen use efficiency ortho-metaQTL in bread wheat unravels concerted cereal genome evolution. *Plant J* 65(5):745–756
- Ramakrishna W, Dubcovsky J, Park YJ et al (2002a) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. *Genetics* 162:1389–1400
- Ramakrishna W, Emberton J, Ogden M et al (2002b) Structural analysis of the maize *Rp1* complex reveals numerous sites and unexpected mechanisms of local rearrangement. *Plant Cell* 14:3213–3223
- Rieseberg LH, Willis JH (2007) Plant speciation. *Science*. 317(5840):910–914
- Ronaghi M, Uhlén M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science* 281(5375):363–365
- Rong J, Feltus FA, Liu L et al (2010) Gene copy number evolution during tetraploid cotton radiation. *Heredity*. 105(5):463–472

- Rosloski SM, Jali SS, Balasubramanian S et al (2010) Natural diversity in flowering responses of *Arabidopsis thaliana* caused by variation in a tandem gene array. *Genetics* 186(1):263–276
- Rouard M, Guignon V, Aluome C et al (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res* 39(Database issue):D1095–102
- Salse J, Feuillet C (2007) Comparative genomics of cereals. In: Varshney R, Tuberosa R (eds) *Genomics-assisted crop improvement*. Springer-Verlag, Berlin, pp 177
- Salse J, Bolot S, Throude M et al (2008a) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 20:11–24
- Salse J, Chagué V, Bolot S et al (2008b) New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative *Aegilops speltoides*. *BMC Genomics* 9:555
- Salse J, Abrouk M, Murat F et al (2009a) Improved standards and new comparative genomics tools provide new insights into grasses paleogenomics. *Brief Bioinform* 10:619–630
- Salse J, Abrouk M, Bolot S et al (2009b) Reconstruction of monocotyledonous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci U S A*. 106:14908–14913
- Salse J, Feuillet C (2011) Paleogenomics in cereals. *Compte rendus de l'Académie des Sciences*. 334:205–211
- Salse J (2012) *In silico* archeogenomics unveils modern plant genome organisation, regulation and evolution. *Curr Opin Plant Biol* 15(2):122–130
- Salvi S, Sponza G, Morgante M et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Natl Acad Sci U S A*. 104(27):11376–11381
- Sankoff D, Zheng C, Zhu Q (2010) The collapse of gene complement following whole genome duplication. *BMC Genomics* 11:313
- Sato S, Nakamura Y, Kaneko T et al (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res* 15(4):227–239
- Satoh-Nagasawa N, Nagasawa N, Malcomber S et al (2006) A trehalose metabolic enzyme controls inflorescence architecture in maize. *Nature* 441(7090):227–230
- Springer NM, Ying K, Fu Y et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 5(11):e1000734
- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112–1115
- Schnable JC, Springer NM, Freeling M (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc Natl Acad Sci U S A*. 108(10):4069–4074
- Schnable JC, Wang X, Pires JC et al (2012) Escape from preferential retention following repeated whole genome duplications in plants. *Front Plant Sci* 3:94
- Scherrer B, Isidore E, Klein P et al (2005) Large intraspecific haplotype variability at the *Rph7* locus results from rapid and recent divergence in the barley genome. *Plant Cell* 17:361–374
- Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278):178–183
- Schubert I and Lysak MA (2011) Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet* 27(6):207–216
- Shulaev V, Sargent DJ, Crowhurst RN et al (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* 43(2):109–116
- Sinha AU, Meller J (2007) Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics* 8:82
- Smith LM, Bomblies K, Weigel D (2011) Complex evolutionary events at a tandem cluster of *Arabidopsis thaliana* genes resulting in a single-locus genetic incompatibility. *PLoS Genet* 7(7):e1002164
- Stein N, Feuillet C, Wicker T et al (2000) Subgenome chromosome walking in wheat: a 450-kb physical contig in *Triticum monococcum* L. spans the *Lr10* resistance locus in hexaploid wheat (*Triticum aestivum* L.). *Proc Natl Acad Sci USA*. 97:13436–13441

- Stein N, Prasad M, Scholz U et al (2007) A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor Appl Genet* 114(5):823–39
- Sutton T, Whitford R, Baumann U et al (2003) The *Ph2* pairing homoeologous locus of wheat (*Triticum aestivum*): identification of candidate meiotic genes using a comparative genetics approach. *Plant J* 36(4):443–456
- Tang H, Wang X, Bowers JE et al (2008a) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 18(12):1944–1954
- Tang H, Bowers JE, Wang X et al (2008b) Synteny and collinearity in plant genomes. *Science* 320:486–488
- Tang H, Bowers JE, Wang X et al (2010) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. *Proc Natl Acad Sci U S A*. 107(1):472–477
- Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res* 16(7):934–946
- Throude M, Bolot S, Bosio M et al (2009) Structure and expression analysis of rice paleoduplications. *Nucleic Acids Res* 2009:37:1248–1259
- Turner A, Beales J, Faure S et al (2005) The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Science* 310:1031–1034
- Tuskan GA, Difazio S, Jansson S et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313(5793):1596–1604
- Tyrka M, Perovic D, Wardynska A, Ordon F (2008) A new diagnostic SSR marker for selection of the Rym4/Rym5 locus in barley breeding. *J Appl Genet* 49(2):127–134
- Van de Peer Y, Fawcett JA, Proost S et al (2009a) The flowering world: a tale of duplications. *Trends Plant Sci* 14(12):680–688
- Van de Peer Y, Maere S, Meyer A (2009b) The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10(10):725–732
- Vandepoele K, Saey S, Simillion C et al (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcollinearity between Arabidopsis and rice. *Genome Res* 12:1792–1801
- Vandepoele K, Simillion C, Van de Peer Y (2003) Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15:2192–2202
- Velasco R, Zharkikh A, Affourtit J et al (2010) The genome of the domesticated apple (*Malus × domestica* Borkh.). *Nat Genet* 42(10):833–839
- Wang J, Long Y, Wu B et al (2009a) The evolution of brassica napus FLOWERING LOCUS T paralogs in the context of inverted chromosomal duplication blocks. *BMC Evol Biol* 9:271
- Wang X, Gowik U, Tang H et al (2009b) Comparative genomic analysis of C4 photosynthetic pathway evolution in grasses. *Genome Biol* 10(6):R68
- Weber JL, Myers EW (1997) Human whole-genome shotgun sequencing. *Genome Res* 7(5):401–409
- Wicker T, Yahiaoui N, Guyot R et al (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell* 15:1186–1197
- Wicker T, Yahiaoui N, Keller B (2007) Contrasting rates of evolution in pm3 Loci from three wheat species and rice. *Genetics* 177(2):1207–1216
- Woodhouse MR, Schnable JC, Pedersen BS et al (2010) Following tetraploidy in maize, a short deletion mechanism removed genes preferentially from one of the two homologs. *PLoS Biol* 8(6):e1000409
- Xiao H, Jiang N, Schaffner E et al (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*. 319(5869):1527–1530
- Yahiaoui N, Srichumpa P, Dudler R et al (2004) Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat. *Plant J* 37:528–38
- Yan L, Loukoianov A, Tranquilli G et al (2003) Positional cloning of the wheat vernalization gene *VRN1*. *Proc Natl Acad Sci U S A*. 100:6263–6268

- Yan L, Loukoianov A, Blechl A et al (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* 303:1640–1644
- Yan L, Fu D, Li C, Blechl A et al (2006) The wheat and barley vernalization gene *VRN3* is an orthologue of *FT*. *Proc Natl Acad Sci U S A*. 103(51):19581–19586
- Zhang X, Wang L, Yuan Y et al (2011) Rapid copy number expansion and recent recruitment of domains in S-receptor kinase-like genes contribute to the origin of self-incompatibility. *FEBS J* doi:10.1111/j.1742-4658.2011.08349.x
- Zhu H, Han X, Lv J et al (2011). Structure, expression differentiation and evolution of duplicated fiber developmental genes in *Gossypium barbadense* and *G. hirsutum*. *BMC Plant Biol* 11:40

Chapter 8

Non-invasive Phenotyping Methodologies Enable the Accurate Characterization of Growth and Performance of Shoots and Roots

Marcus Jansen, Francisco Pinto, Kerstin A. Nagel, Dagmar van Dusschoten, Fabio Fiorani, Uwe Rascher, Heike U. Schneider, Achim Walter and Ulrich Schurr

Contents

8.1 A Growing Number of Imaging Applications Enrich the Plant Phenotyping Portfolio	174
8.2 Precision Phenotyping of Canopies Structure and Photosynthetic Performance	178
8.3 Non-invasive Fluorescence Imaging of Arabidopsis Enables the Quantification of Phenotypic Diversity Driven by Genetic and Environmental Factors	184
8.4 Nuclear Magnetic Resonance Imaging (MRI): A Tool for Characterizing and Optimizing the Dynamic Processes of Rhizogenesis and Root Growth of Cuttings ...	191
8.5 Conclusions	201
References	202

Abstract Significant improvements of the resource-use efficiency of major crops are required to meet the growing demand of food and feed in the next decades in a sustainable way. Breeding for new varieties and modern crop management aims at obtaining higher and more stable yields by optimizing plant structure and function under different environmental conditions. The development and application of non-invasive methods to estimate plant parameters underlying heritable traits are key enabling components. To address this demand, recently an increasing number of imaging technologies have started to be applied in plant research to analyze various types of genotype collections. Some of these applications are mature and suitable to be scaled-up to higher throughput; others require validation beyond proof-of-concept. In this chapter firstly we present an overview of available methods while stressing the current limitations to be taken into account for correct

M. Jansen (✉) · F. Pinto · K. A. Nagel · D. van Dusschoten · F. Fiorani · U. Rascher · H. U. Schneider · A. Walter · U. Schurr
Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum, Jülich GmbH, 52425 Jülich, Germany
email: m.jansen@fz-juelich.de

A. Walter
Institute of Agricultural Sciences, ETH Zürich, Universitätstrasse 2,
8092 Zürich, Switzerland

interpretation of the results. Secondly, we focus on three different case studies by our lab demonstrating the applicability of multispectral, fluorescence, and magnetic resonance imaging for various research questions applicable to controlled environments and to the field. Taken together, these case studies highlight that a variety of non-invasive plant phenotyping methods are essential tools not only for functional genomics, but also for plant selection and breeding. In addition, these experiments underline the need of developing methods tailored to different plant species and at various cultivation systems and scales.

Keywords Plant phenotyping · Non-invasive imaging · Chlorophyll fluorescence · Hyperspectral imaging · Nuclear magnetic resonance imaging (MRI) · Vegetation index

8.1 A Growing Number of Imaging Applications Enrich the Plant Phenotyping Portfolio

Plant phenotypes are dynamic and arise from the complex interaction of genetically encoded molecular networks with multiple environmental factors to which the plant is exposed simultaneously (Walter and Schurr 2005; Walter et al. 2009). In addition, plant responses to the environment are often cumulative and observed phenotypes at a given developmental stage are the result of individual life history. Plant phenotyping however still relies in many cases on traditional methods, e.g. manual measurements or visual estimations (e.g., in the field). To contribute to the selection of genotypes characterized by higher and more stable yields, plant scientists need to tackle this complexity by increased integration of molecular-mechanistic knowledge with accurate measurements of plant performance (Passioura 2010). Also, any technology should be embedded in experimental matrices addressing major factors or their combinations that are relevant for the evaluation of field data (Mittler and Blumwald 2010).

Several imaging methods integrated with appropriate, controlled indoor cultivation systems or by direct deployment of sensors at the field scale offer possible solutions for different research questions applied to shoots and roots (Table 8.1). In this section we first summarize the methodologies that are useful for studying dynamics of whole shoot development and photosynthesis including the evaluation of canopy structure and photosynthesis in the field; secondly we briefly introduce recent advances in root imaging with particular focus on Magnetic Resonance Imaging (MRI).

Automated prototypes using 2D RGB (red-green-blue) or fluorescence imaging have been designed specifically for the extraction of shoot parameters of *Arabidopsis thaliana* (Granier et al. 2006; Walter et al. 2007; Jansen et al. 2009; Arvidsson et al. 2011; Skirycz et al. 2011) or small stature cereals (Berger et al. 2010). These automated systems are suitable for screening genotype panels for variability in (projected) leaf area and developmental dynamics in detailed time course experiments under non-limiting and limiting growth conditions. In addition, shape and geometry of the

Table 8.1 The number of phenotypic parameters that can accurately be measured with imaging techniques is increasing allowing higher throughput both in climate-controlled environments and in the field. We list commonly used and specific imaging methodologies together with recent references (reviews or other significant publications). In addition to the imaging methods, tomographic methods such as Positron Emission Tomography (PET; Jahnke et al. 2009), X-ray Computed Tomography (CT; e.g., Gregory et al. 2003) and neutron tomography (Moradi et al. 2011) are actively being developed for plant roots, and optical scanning methods, such as Laser Induced Fluorescence Transients (LIFT; Kolber et al. 2005), for canopy photosynthesis

Plant parameters	Imaging methods	Experimental setups
Projected shoot area (correlation with biomass); greenness; 2D shoot and root geometry	Color imaging in the visible range (Arvidsson et al. 2011)	Controlled environment
Physiological status of photosystems (PS II)	Chlorophyll fluorescence (Jansen et al. 2009; Rascher et al. 2009)	Controlled environment; field
Canopy temperature	Thermal imaging (Munns et al. 2010)	Controlled environment; field
Various pigment content and canopy properties, such as leaf area index	Imaging spectroscopy/hyperspectral imaging (Ustin and Gamon 2010; Malenovsky et al. 2009)	Field; controlled environment
Structural features in 2D or 3D; water content, diffusion and flow; distribution of specific chemical compounds such as sugars or lipids if highly concentrated	Magnetic Resonance Imaging (MRI) (Simpson et al. 2011; van As 2007; van As et al. 2009)	Controlled environment

shoot is calculated by image analysis as well as photosystems' physiological status. Taken together, these methodologies provide proxies for shoot biomass development (Golzarjan et al. 2011) and for the evaluation of responses of the photosynthetic machinery (photosystems and electron transfer) to environmental challenges in controlled or semi-controlled environments (Jansen et al. 2009). However, we consider that applications of fluorescence imaging in controlled environments allowing meaningful interpretation of results are still lacking specifically for non-rosette plants (see also considerations in Berger et al. 2010; Furbank and Tester 2011). Nonetheless, several fluorescence imaging systems are commercially available and, at least for individual leaves, responses to biotic (Osmond et al. 1998) and abiotic triggers (Walter et al. 2004) can be studied in detail. In the field it also became feasible recently to detect sun-induced fluorescence using the Fraunhofer Line Depth (FLD) detection principle (Moya et al. 2004; Rascher et al. 2009).

In addition to visible and fluorescence imaging, thermal imaging is increasingly used to map canopy temperature and for establishing screening protocols to identify in particular genotypes with high water use efficiency (Munns et al. 2010). Although thermography is an attractive technique, we stress that result interpretation and its use as a proxy for transpiration requires careful experimental calibration and a deep understanding of the physical principles of heat exchange at the leaf surface (Kümmerlen et al. 1999).

Recently, spectral analysis approaches have arisen as a versatile and accessible tool for non-destructive observation of vegetation, and sensors detecting spectral distribution of the radiation reflected by vegetation are becoming affordable and more widely used. The radiative properties of plant leaves or canopies can be used for determining structure and physiological status of the vegetation. The portion of radiation that is reflected, absorbed and transmitted for a specific wavelength at the leaf or canopy scale is determined by: (i) leaf structure and chemical composition; (ii) optical properties of the canopy and soil; and (iii) external effects like illumination and the observation geometry, which is defined by the position of the sun, vegetation and the sensor (Goel 1988, 1989; Chen et al. 2000). Optical spectroscopy uses mainly the reflected part of the radiation to retrieve information about biochemical and structural properties of vegetation. For instance, the spectral reflectance of vegetation is characterized by a low reflectivity in the visible part of the spectrum (400–700 nm) due to a strong absorption by photosynthetic pigments, while a high reflectivity in the near infrared (700–1100 nm) is produced by a high scattering of light by the leaf mesophyll tissues (Knippling 1970; Rascher et al. 2010). In addition, in the shortwave infrared part of the spectrum (1100–2500 nm) the reflectance intensity is affected by the water content of plant tissues (Danson et al. 1992, Rollin and Milton, 1998, Rascher et al. 2010).

The characteristics of optical spectroscopy make this technique highly suitable for fast, non-invasive and reproducible measurements on plant function. The addition of spatial information by imaging spectroscopy offers new opportunities for plant phenotyping. The data can be seen as a cube, with the X and Y axes corresponding to two spatial dimensions and the Z axis to the spectrum (Fig. 8.1). For each pixel, a continuous spectrum of reflectance is obtained. The interpretation of the data is not always simple as the spectral reflectance is strongly affected by the spatial distribution of the different elements involved in the interaction between radiation and vegetation (Goel 1988; Myneni et al. 1989; Barton and North 2001).

Research on the dynamics of root growth requires different methodological approaches compared to research on shoots. This field of research experienced substantial improvements in recent years because of a series of technological developments. Rhizotron cultivation with protocols using artificial substrates such as agar and soil mixtures, as well as natural soils, was combined with different techniques visualizing root systems, such as cameras (Armengaud et al. 2009; Hargreaves et al. 2009; Nagel et al. 2009, 2012; Rascher et al. 2011) or neutron radiography (Carminati et al. 2010). Combination with appropriate analytical tools yields 2D data sets of root systems suitable for calculating features of root system architecture such as root diameter, root length, density, and branching angles (Hargreaves et al. 2009; Nagel et al. 2009, 2012; Rascher et al. 2011). For visualization of roots in three-dimensional container systems protocols have been developed for plant cultivation in transparent gellan gum for image acquisition with cameras (Iyer-Pascuzzi et al. 2010; Clark et al. 2011). For opaque systems, i.e. roots in soil, substantial efforts have been made to implement suitable measuring protocols with X-ray computed tomography (CT; Heeraman et al. 1997; Gregory et al. 2003; Pierret et al. 2003; Hargreaves et al. 2009), neutron tomography (Moradi et al. 2011) and nuclear magnetic

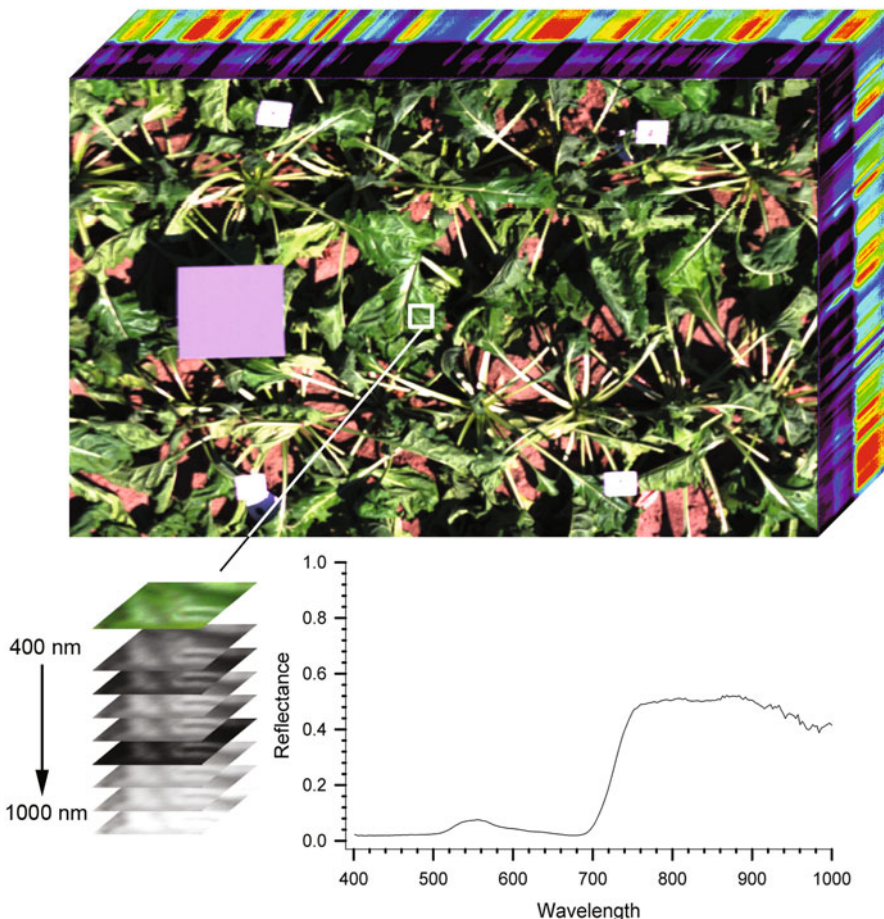


Fig. 8.1 Scheme of 3-dimensional data cube from a sugar beet canopy obtained by imaging spectroscopy. The hyperspectral data cube was taken from four meters above the canopy in the experimental field site Klein-Altendorf of the University of Bonn, Germany. Two dimensions account for the spatial information, while the third dimension codes for the spectral information. Each pixel in a scene is associated with a continuous spectrum of relative reflectance. In the left part of the image, the gray reflectance standard is visible. The stack (*lower left corner*) exemplifies the intensity of some spectral bands between 400 and 1000 nm, taken from the region of interest. The reflectance spectrum (*lower right side*) gives the hyperspectral reflectance values (relative to incoming radiation) of this region of interest

resonance imaging (MRI; e.g. Bottomley et al. 1986, 1993; Brown et al. 1991; Menzel et al. 2007; Jahnke et al. 2009; Nagel et al. 2009; Rascher et al. 2011). A recent summary of achievements in research on root system architecture using CT is given by Zhu et al. (2011). Here we will focus on describing the potential of MRI for root phenotyping.

In brief, MRI detects signals from protons (hydrogen nuclei) in a given specimen. The signals are measured in a spatial array and plotted as an image, where the brightness of the pixels is roughly proportional to the proton density. Though being

particularly valuable for investigations on plant organs in opaque media such as soil, application of MRI for studies of entire root systems in soil is still very scarce and challenging and, thus, far from being a routine approach for phenotyping. However, when ferromagnetic particles are removed and proper soil types and soil moisture are chosen, decent MRI images can be acquired (e.g. Rogers and Bottomley 1987; Brown et al. 1991; Menzel et al. 2007; Hillnhütter et al. 2012; Rascher et al. 2011). A very good case study of Brown et al. (1991) showed that MRI with special soil mixtures can even allow for accurate non-invasive quantification of root length during growth. So far, most MRI studies on root systems did not focus much on measuring speed. Rather, to achieve high-quality images they used comparably long measuring times per data set ranging from tens of minutes up to hours (Brown et al. 1991; Hillnhütter et al. 2012). However, development of fast measuring and data evaluation protocols as shown for other MRI applications (e.g. Meininger et al. 1997; Rokitta et al. 1999), in combination with automation of sample handling, would allow for the use of MRI for moderate-throughput scanning of root systems for phenotyping purposes, similar to its employment in food and materials sciences.

Each phenotyping method opens a specific window to an aspect of the multitude of phenotypic properties that make up structure and functionality of a whole plant. In the following sections we present three different case studies that highlight the range of different imaging applications (spectroscopic and fluorescence imaging and MRI) in both controlled environment and field experiments and for different plant species (*Arabidopsis*, barley, petunia and poplar). Taken together, these examples show that selected technologies can be used to address relevant biological questions concerning variability of plant responses to the environment (case studies 1–3) and also be used for advanced pre-selection protocols of valuable breeding populations (case study 3).

8.2 Precision Phenotyping of Canopies Structure and Photosynthetic Performance

In optical spectroscopy of vegetation, numerous analytical procedures exist for the extraction of information from, and interpretation of, reflectance signatures based on the analysis of either continuous spectra or selected spectral regions or wavelengths. The analysis of continuous spectra may be the most complete way but, due to the high number of parameters that can affect spectral features, normally is not a straightforward process. In fact, each wavelength could be considered as a variable. To simplify the analysis, scientists have focused on particular wavelengths that are directly related to reflectance properties of specific molecules or that could represent a proxy to identify stress response at the leaf or canopy level. The combination of reflectance values at two or more specific spectral bands yields so-called vegetation indices (VIs) (Jackson and Huete 1991). In the past decades, a great number of VIs has been developed to quantify (i) pigment contents, (ii) functional and physiological properties, and (iii) structural properties of plants and canopies. VIs offer the following advantages: (i) a small amount of spectral bands is required, simplifying the

Table 8.2 List of the most used vegetation indices. In the formulas R_x denote for relative reflectance at the wavelength or spectral region that is specified in the subscript, i.e. R_{red} codes for the relative reflectance in the red light spectrum and R_{670} codes for reflectance at 670 nm

Vegetation index	Formula
Normalized difference vegetation index	$NDVI = \frac{R_{NIR} - R_{red}}{R_{NIR} + R_{red}}$
Simple ratio	$SR = \frac{R_{NIR}}{R_{red}}$
Enhanced vegetation index	$EVI = 2.5 \frac{R_{NIR} - R_{RED}}{R_{NIR} + 6 \times R_{RED} - 7.5 \times R_{BLUE} + 1}$
Carotenoids reflectance index 1	$CR1 = \frac{1}{R_{510}} - \frac{1}{R_{550}}$
Carotenoids reflectance index 2	$CR2 = \frac{1}{R_{510}} - \frac{1}{R_{700}}$
Anthocyanin reflectance index 1	$AR1 = \frac{1}{R_{550}} - \frac{1}{R_{700}}$
Anthocyanin reflectance index 2	$AR2 = R_{800} \times \left[\frac{1}{R_{550}} - \frac{1}{R_{700}} \right]$
Plant senescence reflectance index	$PSRI = \frac{R_{680} - R_{500}}{R_{750}}$
Photochemical reflectance index	$PRI = \frac{R_{531} - R_{570}}{R_{531} + R_{570}}$

equipment needed for the measurements; (ii) simple calculation; and (iii) reduced use of computing resources. However, by using VIs major portions of the continuous spectrum are not considered and the capability of modern spectrometers is not fully exploited. Thus, VIs are a simple way to quantify plant traits, but may not be sufficient to quantify complex processes that are related to subtle changes within the absorption properties of leaves and canopies, such as the functional reorganization of pigments associated with photosynthetic light protection. Additionally, VIs are known to be greatly affected by the observation geometry and vegetation architecture and may be dependent on the spatial scale of observations and orientation between the plant and the sensor (Blackburn 2007).

Despite the constraints that VIs present, this method is becoming widely adopted among researchers in the area of breeding, precision agriculture and remote sensing, mainly due to its simplicity. Moreover, the introduction of imaging spectroscopy offers a simple way for estimation of functional and structural traits through spatially resolved VIs at leaf and canopy level. A study case is presented in the following, where VIs estimated from hyperspectral images were taken over crop canopies of four barley varieties (Barke, Mauritia, Sebastian, and Wiebke) grown using the standard agricultural management of barley under rainfed conditions.

Measurements were taken at five different dates during the vegetation period, around solar noon (± 2 h). An imaging spectrometer PS V10E (Spectral Imaging Ltd., Finland) was used to obtain the spectrum of vegetation in the visible/ near infrared part of the spectrum (350–1050 nm). Images were acquired at 4 m height in nadir position, each of them covering ground area of 1×1.5 m within each experimental plot. A white panel with Lambertian reflectance was located in each scene and then used as a reference of solar radiation for the reflectance estimation of each pixel within the image.

Table 8.2 shows a selection of VIs that are potentially useful for phenotyping of plant canopies. Traits like Leaf Area Index (LAI), pigment content or even

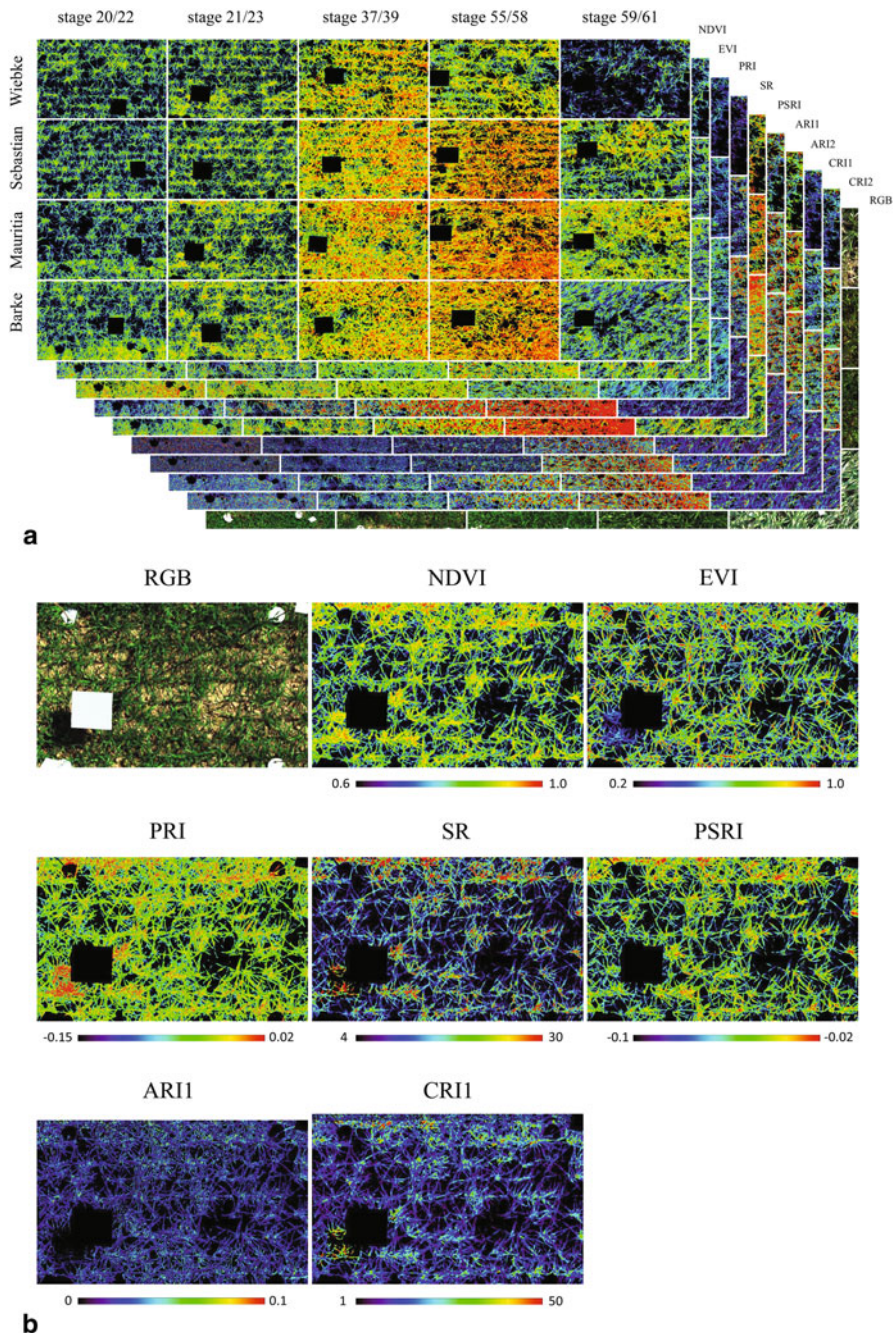


Fig. 8.2 Imaging spectroscopy and the seasonal changes of relevant vegetation indices (VIs) in a barley canopy. VIs were calculated from hyperspectral images in four varieties of barley at five different growth stages according to Zadoks scale (Zadoks et al., 1974) recorded 4 m above the

photosynthetic efficiency have shown correlations with some of these VIs. In this study case, changes in values and spatial distribution through the vegetation period can be observed for most of the VIs (Fig. 8.2a, rows). Additionally, it is also possible to observe some differences between the varieties compared at the same date (Fig. 8.2a, columns). Spatial variations within one image are related to both, differences in plant function and differences in illumination conditions determined by the canopy structure.

It has been observed that increases of grain productivity in crops are more related to higher values of vegetative biomass and increases of photosynthesis per unit land area than to higher rates of leaf photosynthesis itself (Richards 2000). In this context, estimation of LAI and biomass are of particular interest to evaluate whether the canopy structure presents optimal conditions for light capture and biomass production. These two parameters have shown good correlation with those VIs based on the differences in reflectance between the red region (RED) and the near infrared (NIR), which is a characteristic feature of green plant tissues (Christensen and Goudriaan 1993; Turner et al. 1999). Two of these indices which are most frequently used are the Normalized Difference Vegetation Index (NDVI) and the Simple Ratio (SR, Table 8.2).

NDVI can be used to track seasonal changes in barley canopies that can be related with structural changes of the canopy or plant cover (Fig. 8.2). For the same reason, it has also been used for indirect estimation of absorbed photosynthetically active radiation (APAR) (Christensen and Goudriaan 1993; Hansen and Schjoerring 2003; Turner et al. 1999).

When observing differences of NDVI within one image (Fig. 8.2b) it can be noticed that in zones with higher density of plants, like in the borders of the images or in the leaves of the upper layer (center of the plant looking from above), values are higher. From the pictures in Fig. 8.2 it becomes obvious that only leaves of the outer, visible canopy contribute to the NDVI image and there is no superimposed information from lower canopy layers.

There is evidence that NDVI correlates with leaf chlorophyll content (Haboudane et al. 2004), meaning that this index can be potentially used to establish phenological changes based in this pigment or observed nitrogen deficit (Hansen and Schjoerring 2003). In our images, seasonal changes observed in NDVI values (Fig. 8.2a, rows) may be closely related with the seasonal changes in chlorophyll content. It can be observed, for example, that at earlier stages, the values of NDVI are more homogeneous and lower than values at the middle of the growing season (Fig. 8.2a). What is more, a dramatic reduction of NDVI is observed towards the last measuring date, indicating a reduction of green material, which coincided with the ripening of the ears and an increased amount of yellow leaves.

Simple Ratio (SR) and Enhanced Vegetation Index (EVI, Table 8.2) were developed in remote sensing science and were shown to give a better estimation at

Fig. 8.2 canopy. White elements are reference targets that are used for calibration. **a** Seasonal variations of NDVI values measured on five growth stages during the vegetation cycle and in comparison of four varieties. **b** Different VIs calculated over a canopy of barley var. Mauritia at growth stage 21/23 (Zadoks scale)

LAI values of dense vegetation. The EVI adds additionally the reflectance value in the blue region to correct the signal from soil background and reduces atmospheric influences including aerosol scattering. However, remote-sensing science normally works with very coarse spatial resolution and single pixels are spectral aggregates over several meters or even kilometers. Consequently, a large number of objects are spectrally mixed in the spectrum for each pixel. Using these VIs in plant and canopy imaging has several disadvantages. EVI, for example, is greatly affected by shadows; significantly lower values appear in shaded areas in comparison to sun illuminated ones (see shadow in lower left corner of EVI image in Fig. 8.2). SR also is affected by shadows (Fig. 8.2), but may provide the advantage to contain also information related to occluded leaf patches in dense canopies (Fig. 8.2b with dense canopy).

Seasonal changes were also observed in VIs, which can be used to estimate other pigments of the leaves (seasonal image data in lower stacks of Fig. 8.2). That is the case of the Carotenoids Reflectance Index 1 and 2 (CRI1 and CRI2, Table 8.2) and the Anthocyanin Reflectance Index 1 and 2 (ARI1 and ARI2, Table 8.2). These indices are based on the strong absorption that these pigments have at 510 nm (carotenoids) and at 550 nm (anthocyanin). Therefore, variations of reflectance values at these specific wavelengths are directly related with the variation in concentrations of these pigments at leaf level. However, chlorophyll also affects the reflectance in this part of the spectrum and therefore is in part eliminated by subtracting reflectance values at wavelengths where chlorophyll just absorbs (700 nm).

Besides chlorophyll, carotenoids are the main pigments of the leaves. Specific structural and physiological functions have been attributed to them such as: energy transfer, participation of light harvest, antioxidants, structural role in photosynthetic membranes, and quenching of chlorophyll excited states (Gitelson et al. 2002). These pigments are located mainly in the vacuoles of epidermal cells and there is evidence that they play an important role in photoprotection by filtering part of the UV radiation and excessive Photosynthetic Active Radiation (PAR). Therefore their biosynthesis may be induced by stresses that lead to a reduction of photosynthetic efficiency, like for example deficiencies in nitrogen, high UV radiation or pathogen infections (Gitelson et al. 2009).

However, only a small number of studies have investigated these indices and their implications at the canopy scale (Blackburn 2007). It must be taken into account that these indices use very specific wavelengths that can be easily confounded by different factors while measuring at this level. Changes in the observation geometry or the presence of shadows can lead to wrong estimations of anthocyanins and carotenoids (Verrelst et al. 2008). A good example can be seen in Fig. 8.2b, where the shadow produced in the lower left corner leads to an underestimation of anthocyanin and an overestimation of carotenoids.

Seasonal changes of these indices, as well as changes of chlorophyll estimated using NDVI, must be interpreted with care. Leaf senescence for example is characterized by a strong reduction of the ratio between chlorophylls and carotenoids. However, in our images, NDVI shows a similar time course as CRI indices. The reason may be structural factors that are increasing values of NDVI or producing an overestimation of the carotenoids indices. Alternatively, the reflectance at

the specific wavelengths used in CRI, as well as for ARI, may have been strongly affected by high content of chlorophyll. Merzlyak et al. (1997) suggest that indices like CRI1 and CRI2 are not suitable for observing leaf senescence because they have shown to be less sensitive at low concentrations of pigments. The Plant Senescence Reflectance Index (PSRI, Table 8.2) was designed to maximize the sensitivity to the ratio of carotenoids to chlorophyll (Merzlyak et al. 1999). As a consequence, an increase in PSRI indicates a decrease of chlorophyll content and the onset of canopy senescence. In this context, the behavior of PSRI is in agreement with the changes observed in CRI1 and CRI2. However, it is still unclear to which magnitude this index can be affected by structural or illumination factors.

Although traits such as LAI and pigment content can work as indirect estimators of the physiological status of the crop, they usually fail when trying to estimate the actual photosynthetic efficiency of the vegetation. So far, the best methodologies for measuring photosynthesis are based on measurements at leaf level, and have succeeded very seldom in up scaling them to canopy or crop. The Photochemical Reflectance Index (PRI), for instance, is a vegetation index designed to detect changes in the xanthophyll pool composition associated with reduction of light use efficiency (Gamon et al. 1990, 1992; Rascher et al. 2007). A decrease in photochemical efficiency via violaxanthin de-epoxidation due to excess of radiation leads to an increase of the absorption at 531 nm, and therefore to lower values of PRI (Table 8.2). It is worthy to note, however, that there is some confusion in the literature since in some studies the formula is inverted and thus depending on specific definition, the PRI may be positively or negatively correlated with photosynthetic efficiency.

In Fig. 8.2, low values of PRI coincide with the part of vegetation which is under the higher light conditions. In contrast, it can be observed that in general higher values of PRI occur where vegetation is in the shadow. Higher values of PRI are also found in the upper part of the canopy. In this case, the vertical orientation of the leaves in these regions has a strong effect on the PRI values, indicating that canopy structure is an important point to consider for a proper interpretation of this index. It has been well established that PRI is strongly affected by the plant/canopy structure and the geometry of the observation (Barton and North 2001; Verrelst et al. 2008), showing large variability even among plants with the same photosynthetic capacity (Guo and Trotter 2004).

PRI decreased strongly in barley canopies at day 188 (growth stage 59-60 according to Zadok scale) of our measurement series, which coincided with a strong increase of anthocyanin, carotenoids and senescence indices, but also with high values of NDVI. This could mean that lower values of PRI at this date indicate a lower photosynthetic efficiency driven by the loss of chlorophyll that can be deduced from the high values of ARI, CRI and PSRI. Alternatively, high NDVI values at this stage of crop development may be explained by a high LAI and biomass. Differences in PRI have been reported to be affected by changes in leaf pigment level (chlorophyll), which would be a reason why PRI is correlated to seasonal changes in light use efficiency (LUE) (Stylinski et al. 2002).

PRI has been studied in numerous analyses using different classes of vegetation from leaf to ecosystem level, at different time scales. Garbulsky et al. (2011) performed a meta-analysis based on more than 80 publications about PRI, and concluded that in general, this index shows a good correlation with different physiological variables at different time and spatial scales, especially with effective quantum yield ($\Delta F/F_m'$) and LUE. At the canopy level, the good estimation of LUE by PRI suggests that optical properties of the upper canopy region can be used to estimate the photosynthetic efficiency of the whole canopy (Garbulsky et al. 2011).

As shown above, drawing conclusions derived from VIs is often complex. A main issue at canopy scale is its complex 3D structure, in which the reflected radiation is highly determined not only by the external light conditions but also by its architecture. In this regard, imaging spectroscopy on the canopy scale offers the unique advantage to have the spatial information that allows identifying specific elements within the canopy or specific regions under different light conditions. Moreover, its combination with methods for 3D canopy reconstruction can be a powerful tool for a better understanding of the light-canopy-sensor interaction (Rascher et al. 2011) and provide information how spectral reflectance and VIs are affected by the canopy structure.

Although VIs are an easy and fast way to obtain information on some traits in vegetation, they can greatly be influenced by other factors and thus their robustness is limited. Thus, quantitative conclusions have to be drawn with care. It is likely that vegetation properties or physiological changes rather affect larger regions of the spectrum than narrow bands. Analyses of continuous spectra are promising, but imply the use of complex statistics and modeling. Some of the current and most powerful methodologies are Partial Least Square Regression (PLSR; Feilhauer et al. 2010), supervised and un-supervised endmember selection and unmixing, continuous support vector machines (Hostert et al. 2005), multi-block analysis (Eiden et al. 2007) or simplex volume maximization (Römer et al. 2012). All these methods were shown to provide significantly more accurate results than the use of VIs; however, their application for phenotyping purposes requires major adaptation of computer algorithms and data processing. We nevertheless expect that such advanced methods may become more widely used in the future. However, to date, they remain restricted to a few case studies.

8.3 Non-invasive Fluorescence Imaging of Arabidopsis Enables the Quantification of Phenotypic Diversity Driven by Genetic and Environmental Factors

Besides the selection of appropriate measurement sensors and indicators, minimizing unintended environmental influences on phenotypes by standardization of procedures and randomization of plant layouts is crucial to diminish the influence of external factors on the experimental outcome (Hurlbert 1984; Schilling et al. 2008). By

developing and implementing such standard operation protocols (SOP) of Good Phenotyping Practice (GPP), phenotypic variability can be assigned to different genetic properties of ecotypes, or to environmental factors that were tested, respectively. Such genetic and phenotypic diversity of *A. thaliana* can be exemplified by comparing phenotypes of plants grown in different environments and of ecotypes originating from various geographical locations.

In the case study presented here we set out to validate our *Arabidopsis* screening platform, GROWSCREEN FLUORO (Jansen et al. 2009) to establish detailed SOPs including the management of growth conditions and analyze the sensitivity of commonly used ecotypes to variable levels of irradiance, water supply, and nutrients. *A. thaliana* plants were grown in a growth chamber at controlled non-limiting environmental conditions while ecotypes and other environmental conditions were varied systematically.

Growth as well as shoot architecture responded strongly when one specific ecotype (Col-0) was exposed to different water and light regimes (experiment 1; Fig. 8.3a–d) or different types of soil substrates (experiment 2; Fig. 8.3e). These responses indicate the strong dependency of the measured traits on environmental conditions. In order to assess phenotypic responses to genomic differences, we grew nine different ecotypes, Be-0, C24, Col-0, Cvi, Eri-1, Hog, Ler, Sha, and Ws-2 in a climate chamber in one set of environmental conditions (experiment 3; Fig. 8.3f). Ecotypes displayed strong differences in size and shape of single leaves and rosettes.

As part of the SOP all plants were treated identically before the analysis and before exposing them to varied environmental conditions to maximize differences in phenotypes: all plants were pre-germinated in soil-filled trays. After cotyledon unfolding, they were transferred to single-plant pots filled with equal volumes of substrate according to Table 8.3 and were watered up to maximum water holding capacity. Thereafter water was withheld until reaching the targeted water holding capacity (Table 8.3) and then the target value was maintained by controlled irrigation. Shoot parameters such as projected leaf area, rosette morphology, and chlorophyll fluorescence were quantified using the screening system GROWSCREEN FLUORO (Jansen et al. 2009).

In experiments 1 and 2, one ecotype, Col-0, was chosen to test its sensitivity to crucial environmental conditions. Typical factors which are varied in laboratory experiments are e.g. growth media, watering intensity and frequency, and light intensity and quality (PAR). Some of these factors were systematically varied to assess their impact (see Table 8.3). With exception of the nearly nutrient-free peat soil (low-N soil), all variable factors were chosen to be in a range that does not create extreme conditions and is congruent to what is described for unstressed growth of *A. thaliana* in many published studies (e.g. Weigel and Glazebrook 2002 or <http://www.hort.purdue.edu/hort/facilities/greenhouse/101exp.shtml>).

Whereas the choice of substrate is a single event at the beginning of the cultivation period, light and water supply are subject to fluctuations or intended changes in experiments in growth chamber or glasshouses during experimental protocols. To assess effects of differences in light and irrigation regimes, Col-0 plants were



Fig. 8.3 *A. thaliana* shoot phenotypes. Sample images were taken at the end of the vegetative growth period in each set. Phenotypic variability of ecotype Col-0 caused by cultivation conditions (a–d, experiment 1): moderate light, moist soil (a), low light, moist soil (b), moderate light, wet soil (c), and low light, wet soil (d), or by different growth media (e, experiment 2), or phenotypic variability of different ecotypes (f, experiment 3); scale bar = 10 cm

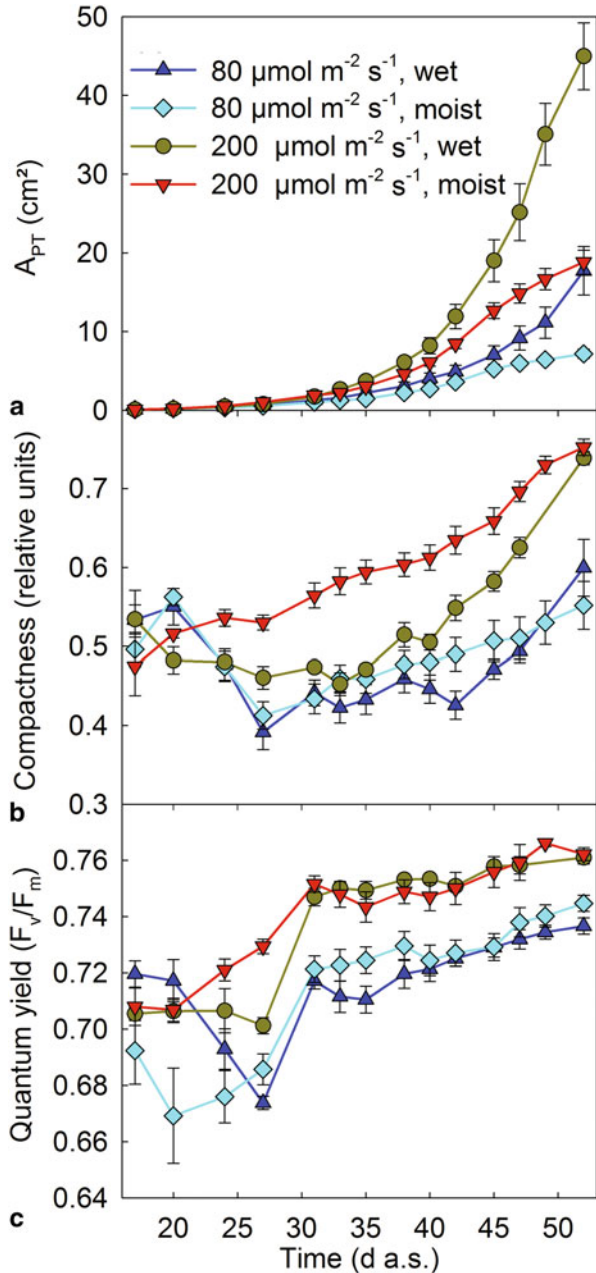
Table 8.3 Phenotypic diversity of *A. thaliana* driven by genetic and environmental factors was exemplary shown by analyzing the ecotype Col-0 under different environmental scenarios (experiment 1: different water and light regimes and experiment 2: different growth media and by comparing ecotypes originating from various geographical locations (experiment 3). All plants were grown under 22°C day and 18°C night temperature, 60 % relative air humidity, and an 8 h/16 h day-night-regime

Exp.	Ecotypes	Soil	Water supply	Light
1	Col-0	Mixed soil: Peat-sand-pumice mixture (NPK 12-12-17)	Moist: 60 % or wet: 95 % of max. water holding capacity	80, 140, or 200 $\mu\text{mol m}^{-2} \text{s}^{-1}$
2	Col-0	Mixed soil (NPK 12-12-17), high-N soil (peat, NPK 25-30-40), intermediate-N soil (peat, NPK 12-14-24) or low-N soil (peat, NPK 5-8-8)	Moist: 60 % of max. water holding capacity	200 $\mu\text{mol m}^{-2} \text{s}^{-1}$
3	Be-0, C24, Col-0, Cvi, Eri-1, Hog, Ler, Sha, Ws-2	Mixed soil (NPK 12-12-17)	Moist: 60 % of max. water-holding capacity	200 $\mu\text{mol m}^{-2} \text{s}^{-1}$

subjected to different light and irrigation regimes (experiment 1). Plants provided with the largest amount of water and highest intensity of light reached the largest final projected leaf area (A_{PT}) of approximately 45 cm² (Fig. 8.4a). In contrast, plants with the lowest supply of water and light reached an A_{PT} of 7 cm² only. Reduction of only one of the two environmental factors, light or water, resulted in plants with similar final leaf area size (18 cm²; Fig. 8.4a). However, the plants reached the final leaf area with pronounced differences in growth rates. At higher light conditions (200 $\mu\text{mol m}^{-2} \text{s}^{-1}$) but reduced water availability (moist soil) plants exhibited initially high growth rates, but leaf expansion rate decreased later, whereas the growth rates of plants grown at low light (80 $\mu\text{mol m}^{-2} \text{s}^{-1}$) and high soil moisture (wet soil) conditions were low at the onset of the experiment and increased at later stages. Besides differences in growth rates, plants were morphologically distinct: plants grown at 200 $\mu\text{mol m}^{-2} \text{s}^{-1}$ in moist soil had shorter petioles—indicated by higher rosette compactness—than all other plant populations (Fig. 8.4b). Plants grown at the same light regime in wet soil treatment increased rosette compactness from 35 d a.s. onwards due to younger leaves growing into spaces between the long-petiole older leaves. With respect to chlorophyll fluorescence all sets of plants had an initial period with variable F_v/F_m . Later on, the efficiency of PSII in plants grown at 200 $\mu\text{mol m}^{-2} \text{s}^{-1}$ stabilized at 0.75, while it stabilized at 0.73 in plants grown at lower light intensities (Fig. 8.4c). The different water availability had minor impact on F_v/F_m .

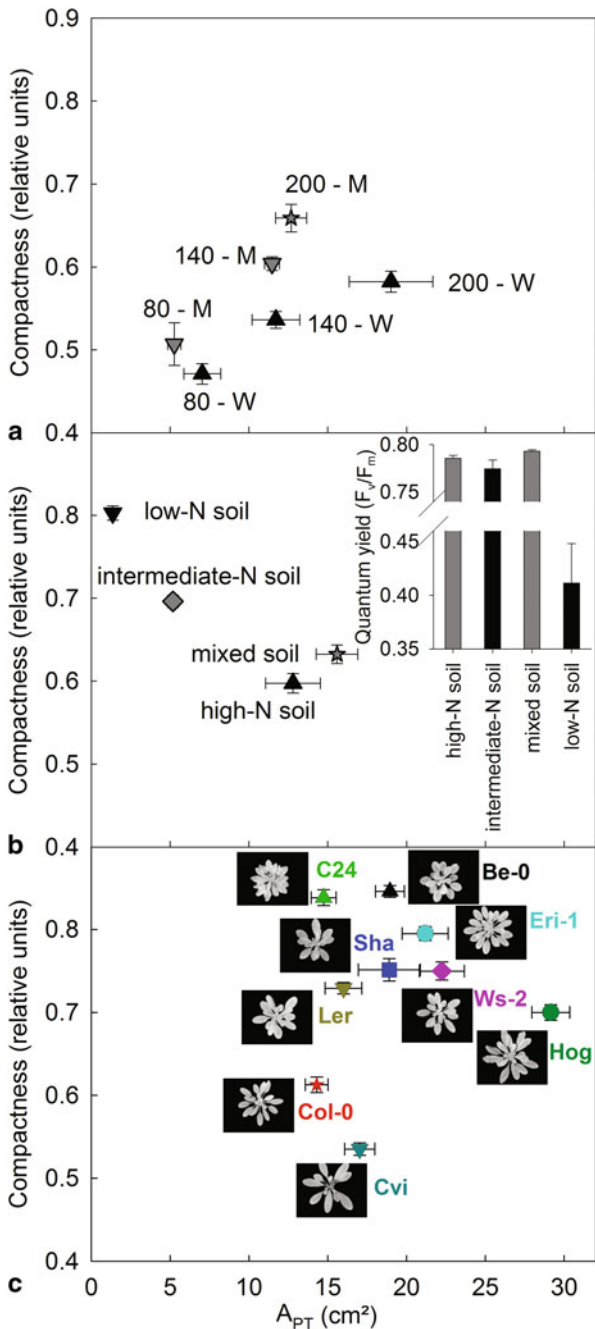
As could be shown, the difference in irrigation of only 10 % (w/w) more water between wet and moist soil treatment and in light intensity of 120 $\mu\text{mol m}^{-2} \text{s}^{-1}$ was sufficient to result in pronounced variations in A_{PT} and rosette shape (Fig. 8.4a, 8.4b). Increased elongation of petioles at low light (Fig. 8.3b, 8.3d and 8.3b) can be interpreted as shade avoidance reaction (Franklin 2008). The results indicate the importance of exactly defined and controlled environmental conditions and cultivation protocols to avoid variability in experimental results and to increase comparability in

Fig. 8.4 Phenotypic variability of *A. thaliana* Col-0 grown in two light regimes (80 or 200 $\mu\text{mol m}^{-2} \text{s}^{-1}$) and two soil moisture conditions (moderately moist and wet, $n = 8$ plants per population, experiment 1): **a** projected leaf area (A_{PT}), **b** rosette compactness, **c** quantum yield of PSII (F_v/F_m); mean values \pm SE



successive and multi-site experiments. A recent inter-laboratory comparison experiment resulted in unexpectedly high variability. However, e.g. light intensity varied between 100 and 180 $\mu\text{mol m}^{-2} \text{s}^{-1}$ in different labs (Massonnet et al. 2010), which may explain a part of the observed variability. The ratios between A_{PT} and rosette compactness of Col-0 plants at different irrigation and illumination levels assessed

Fig. 8.5 Phenotypic variability of *A. thaliana*; ratio between projected leaf area (A_{PT}) and rosette compactness of **a** Col-0 grown at different light regimes (PAR values indicated by numbers) and irrigation levels (experiment 1, 45 d a.s.); $n = 8$ plants per population, **b** Col-0 grown in different soil types (experiment 2, 45 d a.s.); insert: quantum yield of PSII (F_v/F_m) of those plants; $n = 16$ plants per population, **c** ecotypes Be-0, C24, Col-0, Cvi, Eri-1, Hog, Ler, Sha, and Ws-2 at the end of vegetative growth (44 d a.s.); $n = 23$ plants per ecotype (experiment 3). Images show representative individuals of each ecotype; mean values \pm SE



at 45 d a.s. were influenced independently by soil moisture as well as light intensity (Fig. 8.5a). Generally, an increase in soil moisture changed the ratio in a manner that at similar A_{PT} plants in wet soil were less compact. On the other hand, an increase in

light intensity modulated A_{PT} -to-compactness ratio so that plants with comparable A_{PT} were more compact when grown at higher light intensity.

In experiment 2, *A. thaliana* ecotype Col-0 was grown in five potting substrates differing in nutrient composition and soil structure. At 45 d a.s., plants grown in high-N peat soil and intermediate-N peat-sand-pumice mixed soil had similar A_{PT} ranging between 12.8 and 15.6 cm² (Fig. 8.5b), while plants grown in peat soil with either intermediate or low nutrients were significantly smaller (5.2 and 1.4 cm², respectively). Although intermediate-N peat soil and mixed soil had similar nutrient concentrations, plants grown in peat reached only a quarter of the size of plants grown in mixed soil. A potential explanation for these results is that not only the total amount of fertilizer is important to sustain growth, but also the soil structure and thus the accessibility of water and nutrients.

Rosette compactness was also affected by the cultivation substrate, e.g. low fertilizer caused not only smaller sizes but also more compact shoot growth (Fig. 8.5b). Apart from plants grown on almost nutrient-free soil that exhibited a strong decrease of F_v/F_m to 0.4, all plants grown in nutrient richer substrates showed consistent F_v/F_m values of 0.75–0.80 throughout the growth period (Fig. 8.5b, insert). The decrease of low-N plants hints at deficits in PSII functionality at nutrient shortage (Barros and Kuhlbrandt 2009).

At given environmental settings projected leaf area of different *A. thaliana* ecotypes also displayed strong variability (experiment 3, Figs. 8.3f, 8.3c). Differences in A_{PT} developed already early during seedling development (data not shown), which may be due to differences in early seedling vigor or even differences in germination. Later on, similar growth rates in all ecotypes accentuated differences and resulted in considerable diversity in final sizes—the largest ecotype (Hog) reaching a two times larger A_{PT} than the smaller ones (C24 or Col-0; Fig. 8.5c).

Additional strong differences in rosette compactness were found among the ecotypes (Fig. 8.5c). Development of differences in plant size and morphology can be taken as indicators of accumulating impacts throughout the cultivation period determined by ecotype or environment (Fig. 8.5c). Some ecotypes (experiment 3), e.g. Col-0, Ler, and C24 reached similar final A_{PT} , but largely differed in rosette compactness (Fig. 8.5c). Other ecotypes, e.g. Ler, Sha, Ws-2 and Hog, were similar in rosette compactness but differed in A_{PT} .

Intra- and inter-ecotype variability in growth and morphology can be considered as a hallmark of phenotypic plasticity (Reboud et al. 2004). Phenotypic plasticity of a species results from adaptation to natural habitats at the geographic origin and from the ability to cope with environmental stress (Sultan 2000; Koornneef et al. 2004; Pigliucci 2008). In the presented study, we also determined quantum yield of PSII (F_v/F_m), however inter-ecotype differences in F_v/F_m were small throughout the experiment (data not shown).

In summary, phenotypic diversity of *A. thaliana* driven by genetic and environmental factors was clearly shown. Combining phenotypic parameters enables the characterization and quantifying of variability among plant sets in response to genotype and/or environment (Borevitz and Ecker 2004; O'Malley and Ecker 2010). For

the description of a gene function it is necessary to assign the biochemical processes governed by a given gene to physiology and phenotype of the plant (Boyes et al. 2001). Many *A. thaliana* mutants, insertion lines, or activation tagged lines have been investigated to connect phenotypes to genes or genotypes (Bouche and Bouchez 2001; Nakazawa et al. 2003; O'Malley and Ecker 2010). Robust phenotyping methods and exactly designed and recorded protocols (SOPs) are the key to meaningful analyses of gene functions and gene-environment interactions. Moreover, the study shows that it is crucial to precisely control and monitor key environmental factors to which plants are exposed during the experiments. It underlines that—even unwittingly—changed conditions in experimental series with temporal or spatial replications can result in confounding effects.

8.4 Nuclear Magnetic Resonance Imaging (MRI): A Tool for Characterizing and Optimizing the Dynamic Processes of Rhizogenesis and Root Growth of Cuttings

Phenotypic analysis is not only relevant for scientific questions in functional genomics or breeding, but is very valuable in practical plant production. For example vigorous root growth is one phenotypic trait of utmost interest to plant breeders and producers in horticulture. It enables rapid establishment of root systems and, thereby, mechanical support to the plant in a given soil environment, as well as efficient exploration and exploitation of soil resources. Moreover, the potential for rapid development and growth of roots is one essential element of plasticity of the root system to a changing environment. In vegetative multiplication, either for commercial use or for any performance of breeding programs and conservation and restoration programs based on cloning protocols, success notably depends on the phenotypic trait “efficient rooting”, i.e., reliable rhizogenesis and subsequent fast growth of the root system. For plant material produced from cuttings the dynamics of root development is in general not directly deducible from shoot development. Shoot and root growth may show different degrees of coupling during the phase of rhizogenesis and initial root growth (Aminah et al. 1997; Costa and Challa 2002; Kovacevic et al. 2009), depending on transpiration conditions and the complex interrelationship of available carbohydrate sources (leaf area) and sinks (roots, buds) (Dick and Dewar 1992). Thus it is difficult to judge in the initial phase of plant propagation whether cuttings will survive or die. Moreover, certain plant species such as *Populus nigra* and other members of the Salicaceae family can exhibit shoot outgrowth from detached leafless shoot parts without any development of roots. As a result, populations of cuttings of such species are particularly difficult to judge regarding the expected survival rates. In breeding, e.g., of ornamental plants, predictions about survival rates of cuttings can only be made based on longer-term experience gained with the respective plant material. Thus for each and every plant line individual long-term collection of information has to be performed—an economically risky situation for nurseries. Selection of breeding lines based on rooting efficiency, as well as optimization of

propagation protocols for plant material with poor rhizogenesis and root growth are still considerably laborious and time-consuming tasks. Since the 1980s numerous studies have been conducted to develop improved rooting protocols (see e.g. the literature cited by Dick and Dewar 1992, and Costa and Challa 2002). However, these studies focused to a large extent on empirical rather than on causal analysis (but see e.g. Aminah et al. 1997; Costa and Challa 2002; Costa et al. 2007) and mostly determined parameters at the end of the experiments with destructive methods, excavating the belowground part of the cutting and quantitatively evaluating its rooting efficiency after washing off the soil. Systematic non-invasive investigations of the causes of poor rooting efficiency including research on the dynamics of rhizogenesis and root system growth were not possible, but can help accelerating and improving the efficiency of practical plant propagation.

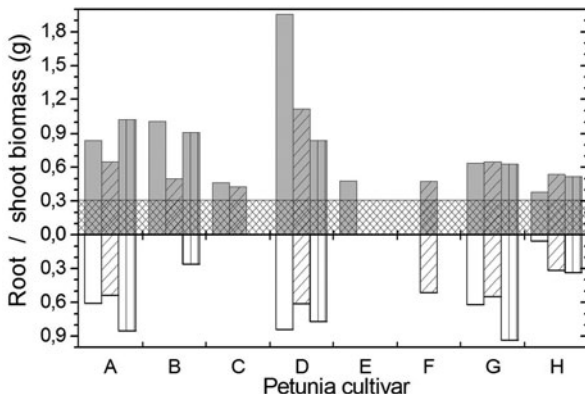
In the following, we describe two case studies to use plant phenotyping for developing practical applications. MRI analysis with short measuring times of less than ten minutes per data set was used to characterize roots in soil, to demonstrate the huge potential of MRI to non-invasively elucidate the dynamics of rhizogenesis and growth of newly formed root systems during cultivation of cuttings for vegetative plant propagation. For these studies we selected petunia because of the economic importance for ornamental plant producers to propagate this species as (cost-) efficiently as possible, and because of the fact that homogeneity and reproducibility of rooting efficiency is decisive in this scenario. In addition, poplar was selected because of the above-mentioned difficulty to predict survival rates and rooting efficiency of cuttings based on shoot development alone.

For the case study on petunia, c. 5 cm long leafy apical cuttings with initial fresh weight of c. 300 mg were prepared for eight different *Petunia x hybrida* cultivars, A–H. The cuttings were individually planted in customary 6-cm diameter pots filled with a 2:1 mixture of sand and natural soil from a field site at Kaldenkirchen, Germany. During a 6-week cultivation period, for three individuals per cultivar shoot development was photographically documented on a daily basis, and MRI measurements were performed on selected dates roughly three, four, five and six weeks after planting to non-invasively check for the root development.

MRI measurements were performed at the NMR facility of IBG-2: Plant Sciences, Forschungszentrum Jülich, Germany, using a 4.7 T Varian VNMRs vertical wide-bore MRI system (Varian Inc., Oxford, UK). We used a so-called spin echo experiment, here with an echo time of 6 ms and a repetition time of 1.5 s. Selection of a field of view (FOV) of $6 \times 6 \text{ cm}^2$ and a matrix of 128×128 pixels resulted in a total measuring time per data set of 3.2 min and an image resolution of $470 \mu\text{m}$. From the image data sets 2D projections over the entire imaged volume were generated after appropriate suppression of the water signals originating from the soil.

During the 6-week observation period for cultivars A, D, G and two out of three individuals of cv. B shoot growth was clearly visible whereas for others, such as C and H, it was hard to decide whether growth took place or not without employing additional technical approaches such as camera-assisted dynamic growth analysis (compare 1.3). Individuals of cultivars E and F occasionally showed yellowing and shedding of leaves towards the end of the observation period. Figure 8.6 gives the root

Fig. 8.6 Biomass, given as fresh weight, of roots (*white columns*) and shoots (*gray columns*) of cuttings of eight petunia cultivars A–H after 6-week cultivation as conventionally determined after harvest. The cross-hatched area marks the average initial biomass of the cuttings at the time of planting. Only data for cuttings without black rot are shown



and shoot biomasses (fresh weight) as determined by conventional harvest after 41-d cultivation. The cross-hatched area in the figure marks the average initial biomass of the cuttings at the time of planting, to facilitate judgment on the amount of shoot biomass newly formed during cultivation. Optical inspection of the below-ground part of all still green cuttings revealed black rot on individuals of cultivars C, E and F, which were discarded from the evaluation. Clearly, the investigated cultivars can be grouped into superior performers with regard to rooting efficiency under the given cultivation conditions, namely A, D and G, and inferior performers B, C, E and H. For the two cultivars C and E none of the cuttings managed to develop roots during the observation period with the given cultivation protocol. For B and F only one individual out of three underwent rhizogenesis.

The reasons were, however, apparently different: Whereas F revealed high susceptibility for black rot infestation all cuttings of B remained healthy, suggesting that their rooting was not sufficiently stimulated by the selected protocol. The highest amount of shoot biomass was newly produced by D and individuals of A and B. As a tendency, higher investment into roots compared to shoots was found for cultivars F, G, H and A, whereas for individuals of D and B less investment into roots compared to shoots had obviously taken place.

Non-invasive investigation of the below-ground part of the cuttings with MRI during the 6-week cultivation period added information on the dynamics of rhizogenesis and root growth for the different cuttings. Figure 8.7 shows exemplary MRI data sets for representative cuttings of cultivars A, B and C 26, 34 and 41 days after planting. The selected MRI protocol allows for the distinction of newly formed adventitious roots from the cuttings and also from the soil. The figure reveals that cutting A showed fastest and most vigorous root development and growth. Roots had emerged after 26 days and almost entirely had reached out into the lower half of the pot by the end of the observation period. Compared to cutting A, rooting of cutting B was considerably delayed: Inspection 26 days after planting revealed no roots, but one week later roots were detectable at a similar amount compared to cutting A at time point 26 d (compare Fig. 8.7, B'' and A'). Further observation of cutting B revealed

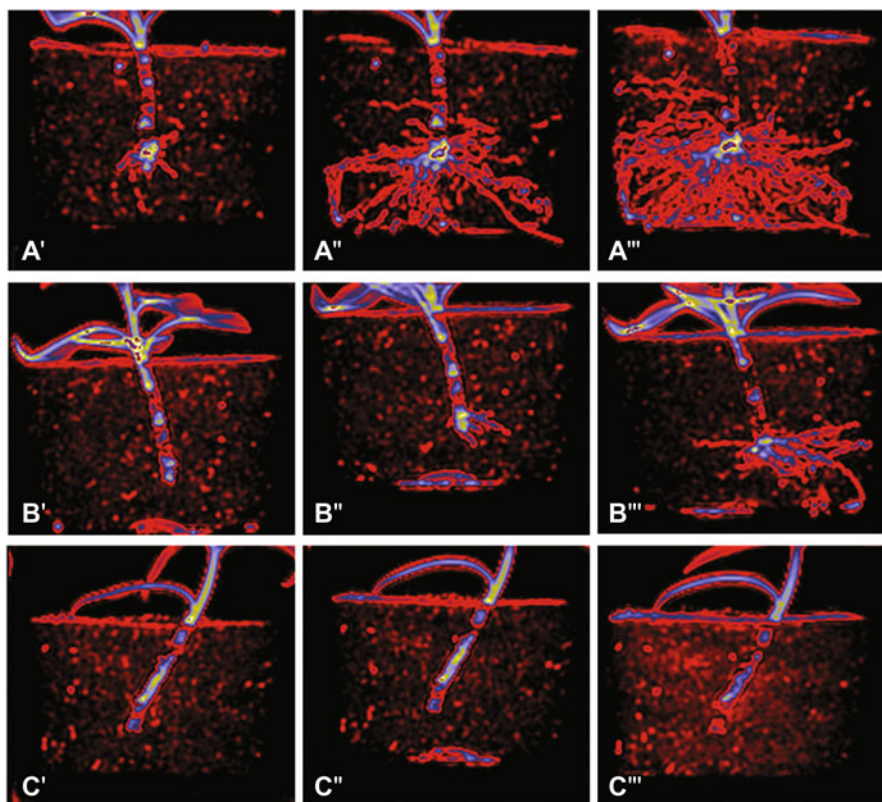


Fig. 8.7 Dynamics of rhizogenesis and progressive root development of individual cuttings of three different petunia cultivars (A, B, C) as revealed by MRI. Measurements were taken 26 d (A', B', C'), 34 d (A'', B'', C'') and 41 d after planting (A''', B''', C'''). Images represent 2D projections of the signals acquired for the entire volume of the specimen. *Black*, no signal; *red*, low signal intensity; *blue*, medium signal intensity; *yellow*, high signal intensity

less root spreading per day compared to cutting A (compare Fig. 8.7, B'' and A'). This suggests that not only rhizogenesis but also root growth was substantially slower in cultivar B compared to cultivar A—calling for an improved cultivation protocol for this cultivar. Possible reasons for this observation could be a slower metabolism or a lower photosynthetic efficiency in B compared to A, or a different partitioning of photoassimilates into roots and shoots for both cultivars. The latter aspect could be investigated e.g. by a combination of gas exchange, MRI and tracer technologies using short-lived ^{11}C and positron emission tomography, as demonstrated by Jahnke et al. (2009).

For the cutting of cultivar C, no root formation was detectable throughout the entire 41-d cultivation period (Fig. 8.7, C'''), suggesting that the cultivation protocol was insufficient also for this petunia cultivar.

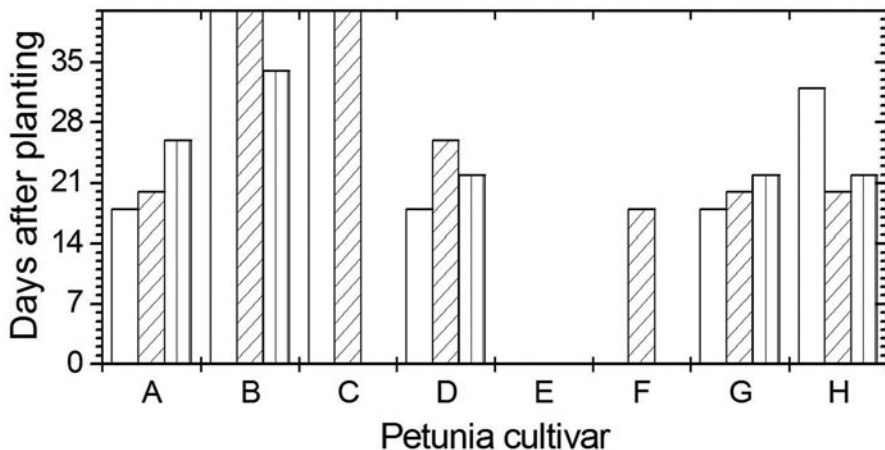


Fig. 8.8 Representation of the dates when roots were first observed with MRI for the individuals of the eight investigated petunia cultivars, A-H. $Y > 41$ d: no root formation during study. Only data for cuttings without black rot are shown

Cutting B exhibited a substantial increase in MRI signal intensity at its basal end after rhizogenesis had occurred (Fig. 8.7, B'') compared to the situation before rhizogenesis (Fig. 8.7, B'). Also cutting A showed particularly intensive signals at its basal end along with vigorous root growth. This is most likely due to a substantial local increase in new biomass development upon meristem activity for adventitious root formation. No such signal increase is observed for the basal end of cutting C (Fig. 8.7, C'-C''). On the contrary, a decrease in MRI signal was observed at this spot at day 41 compared to day 34, obviously due to rotting as revealed by destructive inspection. Evaluation of the data sets for all petunia plants measured confirmed that rhizogenesis was always associated with a local high MRI signal at the base of the cutting (see also Fig. 8.9, H1). This feature could even clearly be identified when the overall image quality was comparably poor and roots could hardly be distinguished from soil water, e.g., in case of excess water (data not shown). Thus this MRI analysis method may help to gain high-accuracy and specific datasets on the onset of root formation.

Figure 8.8 summarizes the dates when roots were first observed by MRI for the investigated petunia specimens. It has to be noted that due to technical reasons, in each of the measurement weeks differences of up to four days could occur between the measurements of the different specimens. Cultivars A, D, G and H showed rooting mostly within two to three weeks. Interestingly, the surviving specimen of F also performed fast with roots observable already at day 18, unlike the only rooted specimen of the slow performer cultivar B with roots first detected at day 34 (Fig. 8.7, B''). Homogeneity of root setting, which is an important parameter for commercial plant production, was highest for cultivar G, while individuals of cultivar H showed a roughly 2-week difference of rooting dates.

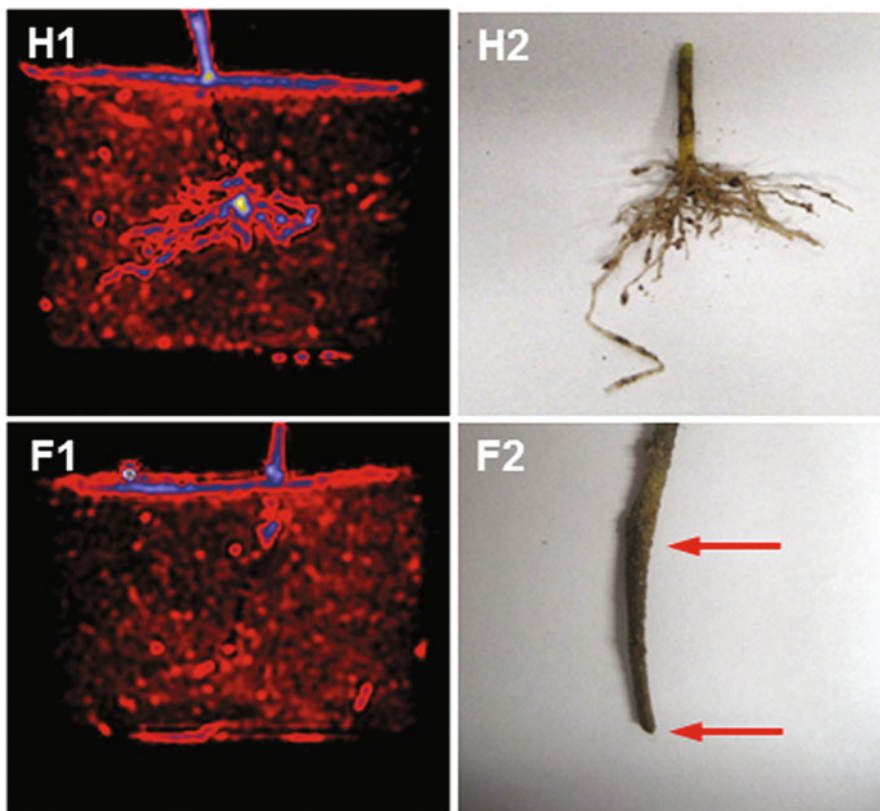


Fig. 8.9 Comparison of root morphologies of individuals from petunia cultivars H and F as revealed by MRI (H1, F1) and photography after excavation (H2, F2) 41 days after planting. The MRI images represent 2D projections of the signals acquired for the entire volume of the specimen. *Black*, no signal; *red*, low signal intensity; *blue*, medium signal intensity; *yellow*, high signal intensity. The *arrows* mark the area of the cutting displaying symptoms of stem black rot

All judgments deduced from MRI whether rooting had taken place or not were confirmed by the results from excavation. Good agreement was also found between the root architecture and morphologies displayed by MRI and the results for the excavated root systems (Fig. 8.9, H1, H2). All cuttings whose basal ends had shown a loss of signal intensity and eventually became indistinguishable from the surrounding soil proved after excavation to be heavily infected by stem black rot (Fig. 8.9, F1, F2).

The good representation of the root morphologies by the MRI 2D projections suggested that calculation of the time courses of root biomass formation for the different petunia cultivars can be achieved on basis of the MRI data. In principle root biomass can be deduced directly from MRI images by multiplying the signal intensity with the pixel volume after calibration against pure water, and by taking into account the usual water content of petunia roots of *c.* 93 % as determined in this study. This was done for single cuttings of the six cultivars which exhibited rooting

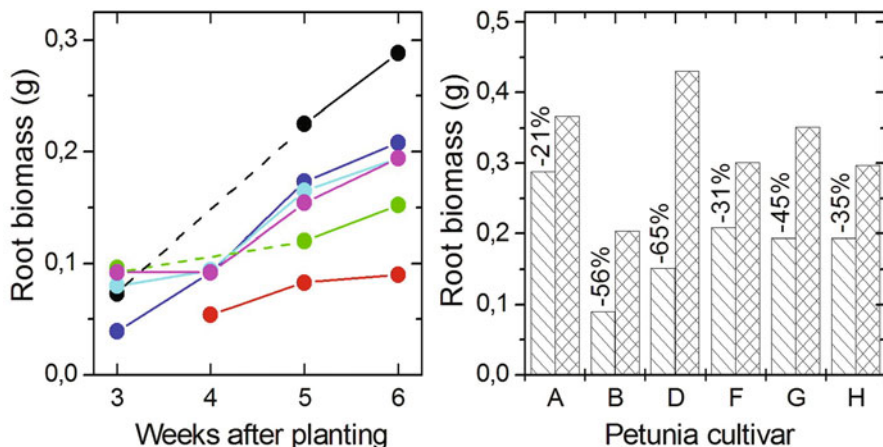


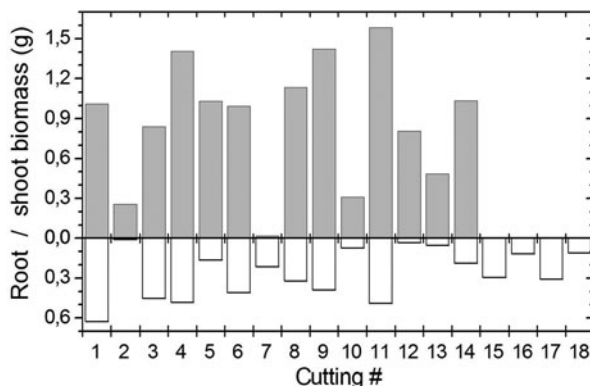
Fig. 8.10 Left: Time course of root biomass development (left) as calculated from MRI 2D projection for individual cuttings from cultivars A (dark blue dot), B (red dot), D (lime green dot), F (azure blue dot), G (aqua dot) and H (pink dot). The dashed lines represent extrapolations since single images exhibited too poor quality for quantitative evaluation. Right: Comparison of root biomass for the same samples as determined by MRI (hatched columns) and conventionally (cross-hatched columns) 41 days after planting. The numbers give the percentage deviation of the MRI results from the real biomass

(Fig. 8.10, left). All time courses show a reasonable increase of root biomass from an initial, comparably low value to the final value during the 41-day observation period. Steepest slopes for root biomass formation (10 mg d^{-1} and 8 mg d^{-1}) were determined for the root systems of the cuttings from cultivars A and F, respectively. Cuttings from B and D exhibited with 2 and 3 mg d^{-1} , respectively, substantially more shallow slopes. Interestingly, the growth of the cuttings from cultivars G and H appeared somewhat slower at the initial compared to later stages, whereas the root system of the cutting from cultivar F showed a more or less constant growth speed throughout the entire observation period. Comparison with the data for the excavated cuttings, however, revealed that the MRI-derived values underestimated the real root biomass at the endpoint of the study by 21–65 % (Fig. 8.10, right).

This pronounced deviation is most likely due to highly discriminating setting of threshold values for subtraction of background water signal: Due to selection of a comparatively high threshold, smaller roots with diameters that are smaller than the edge length of a pixel volume (voxel) are rejected for analysis, thereby causing a considerable reduction of the detected root mass. This effect would get bigger at larger voxel sizes or smaller root diameters.

The reported data suggest that useful MRI images can be obtained for soil-cultivated petunia cuttings within a relatively short measuring time of only 3 minutes and thus can provide a valuable tool for categorizing rooting efficiency of cultivars and for guided improvement of propagation protocols, but also in other phenotyping experiments. The accuracy of the method can be further enhanced by opting for additional averages, at the cost of additional measurement time, and by reducing the

Fig. 8.11 Biomass of newly formed roots (white columns) and shoots (gray columns) of 18 rooted cuttings of cottonwood after 3-week cultivation, given as fresh weight conventionally determined after harvest



soil water content, to facilitate proper setting of the lower threshold for quantitative evaluation.

The second MRI case study comprised 54 6-cm-long leafless cuttings of Eastern cottonwood (*Populus deltoides*) taken from semi-hardwood shoot parts of 26 greenhouse-grown 6-month-old plants. The same type of pots and sand/soil mixture were used as for the petunias, for three-week cultivation. Many of the cuttings showed very fast outgrowth of shoots from buds which had reached heights of up to 2 cm already one week after planting. By the end of the three-week observation period, a high percentage of the shoot-bearing cuttings had formed up to three outgrowths with maximum shoot lengths of 8.5 cm. In total, 41 out of the 54 cuttings (76 %) developed shoots, but three of them died and experienced rotting before the end of the experiment. Additional six cuttings showed symptoms of stem black rot without forming shoots. Five cuttings (9 %) had developed neither shoots nor roots. Only 29 of the shoot-exhibiting cuttings (54 %) had additionally undergone rhizogenesis and root growth. Four individuals had formed roots without forming shoots, thus eventually increasing the percentage of rooted cuttings to 61 % in this study. These results match very well with results from a parallel study on 103 *P. deltoides* cuttings cultivated in commercial propagation substrate under comparable environmental conditions, yielding 71 % and 59 % shoot and root outgrowth, respectively.

Root and shoot biomasses and the respective root/shoot ratios of the cuttings in our MRI study substantially varied between the individuals as shown in Fig. 8.11 for 14 out of the 29 cuttings that exhibited root as well as shoot outgrowth. The root biomass formed by the cuttings which had not developed shoots also considerably varied and occasionally exceeded the values determined for the cuttings with both root and shoot formation (Fig. 8.11).

Consecutive MRI measurements on 24 out of the 54 cuttings two and three weeks after planting added further information regarding the dynamics of root formation and growth of the different cuttings under investigation. The spin echo sequence used featured the same values for echo and repetition times as for the petunias. Selection of FOV = 7.5×7.5 cm² and a 256×256 matrix resulted in a measuring time per data set of 6.5 min and an image resolution of 290 μ m in this study. As shown

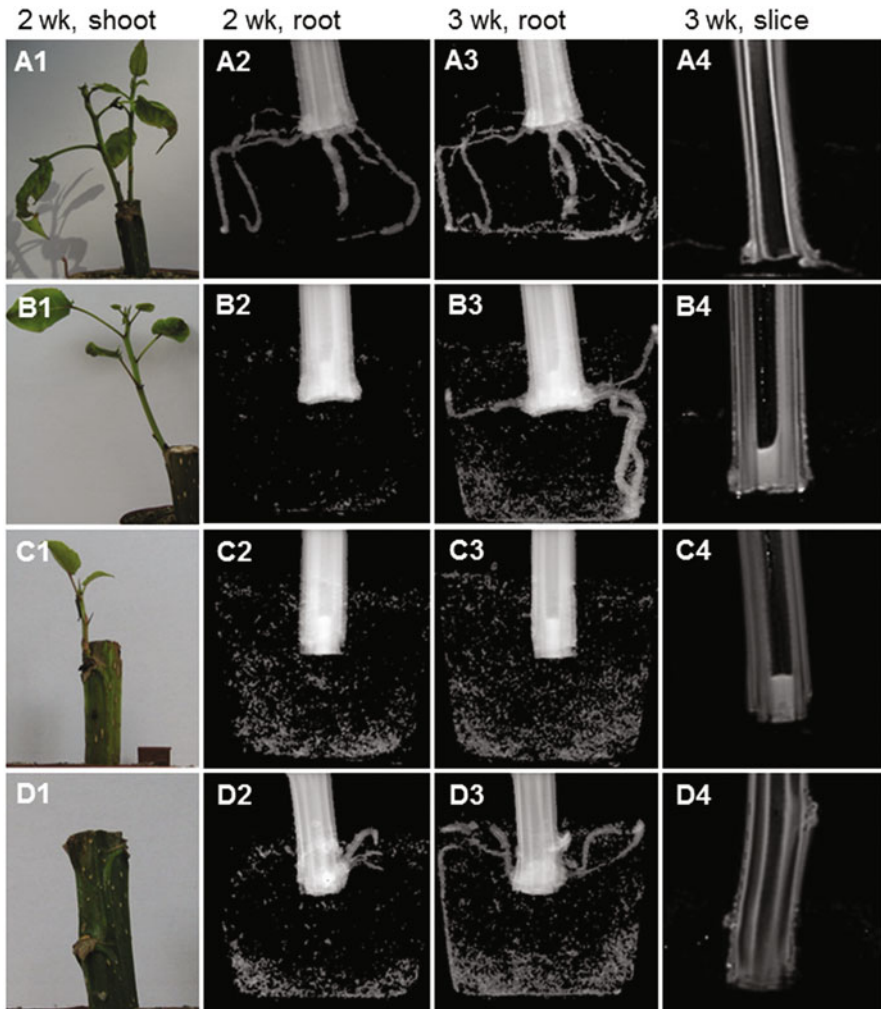


Fig. 8.12 Four examples of shoot development as documented by eye, and root development as documented by 2D MRI projection images, for cuttings of *Populus deltoides*. Two weeks after planting three out of the four cuttings exhibited shoot outgrowth (A1, B1, C1). Roots, however, had developed at the same time only for cuttings A (A2) and D (D2). A rescan of all specimens after 3-week cultivation revealed that cutting B obviously experienced delayed root outgrowth compared to cuttings A and D (see B3) whereas yet no rooting had occurred for cutting C (C3). Images A4, B4, C4 and D4 represent 5-mm-thick sagittal virtual slices of the respective cuttings after 3-week cultivation

in Fig. 8.12 newly formed adventitious roots could clearly be distinguished from the cuttings and also from the soil water by the selected method. 17 out of the 24 cuttings showed rooting. MRI revealed a very good recognition of the four types of outgrowth, namely shoot and root (Fig. 8.12a, 8.12b), shoot but no root (Fig. 8.12c),

no shoot but root (Fig. 8.12d), and no shoot and no root (data not shown). As for the petunias, all results from excavation confirmed the results from MRI about rooting events of the poplar cuttings.

MRI also revealed differences in the dynamics of root formation as shown for the petunias. For eleven out of the 17 rooted cuttings (65%) rooting had already occurred earlier than two weeks after planting (compare Fig. 8.12, A2 and D2). At the time point two weeks after planting the size of the developed root systems varied considerably (Fig. 8.12, A2 and D2). The well-known phenomenon of substantial shoot outgrowth without simultaneous rooting could frequently be documented in this MRI study, as exemplified for cuttings B and C in Fig. 8.12.

More than two weeks after planting, however, many of the respective cuttings eventually showed rhizogenesis. As for the petunias, many poplar cuttings showed particularly high signal intensities at their basal ends (Fig. 8.12, B2/3, C2/3, D2/3). Sagittal virtual slices obtained with MRI of the respective specimens after 3-week cultivation showed that a signal-rich substance had obviously accumulated predominantly in the basal pith region of these specimens (Fig. 8.12, suffices 4). This feature was detected with MRI in 14 cuttings in total, irrespective of being rooted or rootless (Fig. 8.12, B4, C4, D4). On the other hand, it could be absent in cuttings showing particularly vigorous rooting (Fig. 8.12, A4), suggesting that it was not an essential prerequisite for rhizogenesis and/or root growth. Thus, unlike for petunia, particularly high basal MRI signals could not be taken as clear-cut indication for rhizogenesis in the case of the poplar cuttings. Meristems for adventitious root growth, however, could be detected by MRI as smooth, rounded shapes along the otherwise sharp contours of the cuttings (Fig. 8.12, A4, B4, D4). Evaluation of all MRI datasets suggested that cuttings of *P. deltoides* reveal this change of shape obviously substantially earlier compared to the time point when root outgrowth is detectable (Fig. 8.12, B2). By contrast, a cutting which has experienced no rhizogenesis persistently exhibits sharp contours (Fig. 8.12, C3). Thus, the feature of “shape smoothing” may be a valuable trait for optimization of propagation protocols for poplar.

The two presented case studies for petunia and poplar cuttings, though being preliminary due to low sample numbers, show that non-invasive MRI studies could be a valuable tool for the development of phenotyping protocols to optimize vegetative propagation protocols and, thus, the rooting efficiency of cuttings for commercial use, breeding and conservation programs. We have shown that the use of measuring protocols of 3–7 min yielded sufficient image resolution to detect newly formed roots, to elucidate the initial steps of the rooting process of the cuttings and to document the subsequent development of the root systems. In the case of petunia, the study allowed a preliminary categorization of eight cultivars with respect to their rooting efficiency and sensitivity to propagation protocols, based on detection of the time points of root outgrowth and the occurrence of particularly high MRI signal intensities at the base of the cutting. For poplar, indications for the determination of the initial step of root formation based on “shape smoothing” of the basal end of the cutting could be dissected. An explanation for the observed heterogeneity of root and shoot formation could, however, not yet be elaborated.

When using the presented relatively fast measuring protocols the amount of specimens that can be measured on a daily basis with MRI is not so much restricted by the measurement time itself but by the manual positioning of the specimens within the MRI magnet. Recently, we successfully tested a setup with a robot for automated positioning of the specimens in the magnet at the NMR facility of IBG-2: Plant Sciences, Forschungszentrum Jülich, Germany, which in combination with the presented protocols will allow for medium-throughput approaches handling up to *c.* 100 of such small specimens per day. Also, the 2D signal projection provides a time-saving, straightforward approach for data reconstruction and evaluation. More effort, however, has to be put into the development of suitable protocols for proper calculation of root biomass from the 2D signal projections, to allow for additional non-invasive investigations of the dynamic development of root/shoot ratios during the process of rooting. These kinds of measurements can likewise be used to elucidate above- and below-ground dynamics of growth in the presence of biotic and abiotic stresses including nutrient deficiency and drought, and there are many more applications to be addressed with the presented protocols.

8.5 Conclusions

Non-invasive methodologies for the accurate description of plant phenotypes and their dynamic interaction with main environmental factors are invaluable to address the genome to phenome bottleneck (Furbank and Tester 2011) in scientific as well as practical applications. In particular, novel imaging technologies are extensively evaluated by various research groups. Notably, 2D visible and fluorescence imaging are used today for genotype screening. MRI offers unique opportunities for structural-functional analyses especially for studies on the dynamics of root growth and transport in roots. At the field scale, remote sensing of canopy function is constantly increasing its arsenal of sensors and measured parameters and imaging spectroscopy opens a wide range of indices that are related to aboveground traits. We stressed that several technological challenges remain to be addressed to further increase precision, accuracy, robustness and—in some cases—measuring speed and throughput of these methods. In addition, beyond ‘taking pictures’, correct interpretation often requires in-depth understanding of sensor physics and of the biology of the targeted organs. Interestingly, several applications can be used to phenotype a growing number of genetic resources, individually or in combined approaches.

Acknowledgments The authors would like to thank Silvia Braun (case study 2) and Martina Klein, Sabrina Lauter and Carola Mohl (case study 3) for excellent technical assistance. Experiments in case study 1 were supported by Bundesministerium für Bildung und Forschung (BMBF, Germany) in the CROP.SENSE.net consortium. Francisco Pinto was supported by Deutscher Akademischer Austauschdienst (DAAD, Germany) and the Commission for Scientific and Technological Research (CONICYT, Chile). Development of the MRI protocol for petunia in case study 3 was supported by a grant from Bundesministerium für Ernährung, Landwirtschaft und Verbraucherschutz (BMELV, Germany) via Bundesanstalt für Landwirtschaft und Ernährung (BLE, Germany) in the framework of Programm zur Innovationsförderung.

References

- Aminah H, Dick JM, Grace J (1997) Rooting of *Shorea leprosula* stem cuttings decreases with increasing leaf area. *Forest Ecol Manag* 91:247–254
- Armengaud P, Zambaux K, Hills A et al (2009) EZ-rhizo: integrated software for the fast and accurate measurement of root system architecture. *Plant J* 57:945–956
- Arvidsson S, Perez-Rodriguez P, Mueller-Roeber B (2011) A growth phenotyping pipeline for *Arabidopsis thaliana* integrating image analysis and rosette area modeling for robust quantification of genotype effects. *New Phytol* 191:895–907
- Barros T, Kuhlbrandt W (2009) Crystallisation, structure and function of plant light-harvesting complex II. *Biochim Biophys Acta* 1787:753–772
- Barton CVM, North PRJ (2001) Remote sensing of canopy light use efficiency using the photochemical reflectance index—model and sensitivity analysis. *Remote Sens Environ* 78:264–273
- Berger B, Parent B, Tester M (2010) High-throughput shoot imaging to study drought responses. *J Exp Bot* 61:3519–3528
- Blackburn GA (2007) Hyperspectral remote sensing of plant pigments. *J Exp Bot* 58:855–867
- Borevitz JO, Ecker JR (2004) Plant genomics: the third wave. *Annu Rev Genom Hum Genet* 5:443–477
- Bottomley PA, Rogers HH, Foster TH (1986) NMR imaging shows water distribution and transport in plant root systems *in situ*. *P Natl Acad Sci U S A* 83:87–89
- Bottomley PA, Rogers HH, Prior SA (1993) NMR imaging of root water distribution in intact *Vicia faba* L. plants in elevated atmospheric CO₂. *Plant Cell Environ* 16:335–338
- Bouche N, Bouchez D (2001) *Arabidopsis* gene knockout: phenotypes wanted. *Curr Opin Plant Biol* 4:111–117
- Boyes DC, Zayed AM, Ascenzi R et al (2001) Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. *Plant Cell* 13:1499–1510
- Brown DP, Pratum TK, Bledsoe C et al (1991) Noninvasive studies of conifer roots: nuclear magnetic resonance (NMR) imaging of Douglas-fir seedlings. *Can J Forest Res* 21:1559–1566
- Carminati A, Moradi AB, Vetterlein D et al (2010) Dynamics of soil water content in the rhizosphere. *Plant Soil* 332:163–176
- Chen JM, Li X, Nilson T, Strahler A (2000) Recent advances in geometrical optical modelling and its applications. *Remote Sens Rev* 18:227–262
- Christensen S, Goudriaan J (1993) Deriving light interception and biomass from spectral reflectance ratio. *Remote Sens Environ* 43:87–95
- Clark RT, MacCurdy RB, Jung JK et al (2011) Three-dimensional root phenotyping with a novel imaging and software platform. *Plant Physiol* 156:455–465
- Costa JM, Challa H (2002) The effect of the original leaf area on growth of softwood cuttings and planting material of rose. *Sci Hortic* 95(1–2):111–121
- Costa JM, Heuvelink E, Van de Pol PA, Put HMC (2007) Anatomy and morphology of rooting in leafy rose stem cuttings and starch dynamics following severance. *Acta Hortic* 751:495–502
- Danson FM, Steven MD, Malthus TJ, Clark JA (1992) High-spectral resolution data for determining leaf water content. *Int J Rem Sens* 13(3):461–470
- Dick JM, Dewar RC (1992) A mechanistic model of carbohydrate dynamics during adventitious root development of leafy cuttings. *Ann Bot* 70:371–377
- Eiden M, Linden S van der, Schween JH et al (2007) Elucidating physiology of plant mediated exchange processes using airborne hyperspectral reflectance measurements an synopsis with eddy covariance data. In: 10th ISPMRS Conference, March 12–14, 2007, Davos, pp 473–481
- Feilhauer H, Asner GP, Martin RE, Schmidlein S (2010) Brightness-normalized partial least squares regression for hyperspectral data. *J Quant Spectrosc Radiat Transf* 111:1947–1957
- Franklin KA (2008) Shade avoidance. *New Phytol* 179:930–944
- Furbank RT, Tester M (2011) Phenomics—technologies to relieve the phenotyping bottleneck. *Trends Plant Sci* 16:635–644

- Gamon JA, Field CB, Bilger W et al (1990) Remote sensing of the xanthophyll cycle and chlorophyll fluorescence in sunflower leaves and canopies. *Oecologia* 85:1–7
- Gamon JA, Peuelas J, Field CB (1992) A narrow-waveband spectral index that tracks diurnal changes in photosynthetic efficiency. *Remote Sens Environ* 41(1):35–44
- Garbulsky MF, Peuelas J, Gamon J et al (2011) The photochemical reflectance index (PRI) and the remote sensing of leaf, canopy and ecosystem radiation use efficiencies: a review and meta-analysis. *Remote Sens Environ* 115(2):281–297
- Gitelson AA, Zur Y, Chivkunova OB, Merzlyak MN (2002) Assessing carotenoid content in plant leaves with reflectance spectroscopy. *Photochem Photobiol* 75(3):272–281
- Gitelson AA, Chivkunova OB, Merzlyak MN (2009) Nondestructive estimation of anthocyanins and chlorophylls in anthocyanic leaves. *Am J Bot* 96(10):1861–1868
- Goel NS (1988) Models of vegetation canopy reflectance and their use in estimation of biophysical parameters from reflectance data. *Remote Sens Rev* 4:1–122
- Goel NS (1989) Inversion of canopy reflectance models for estimation of biophysical parameters from reflectance data. In: Asrar G (ed) *Theory and applications of optical remote sensing*. Wiley, New York, pp 205–251
- Golzarian MR, Frick RA, Rajendran K et al (2011) Accurate inference of shoot biomass from high-throughput images of cereal plants. *Plant Methods* 7:2
- Granier C, Aguirrezabal L, Chenu K et al (2006) PHENOPSIS, an automated platform for reproducible phenotyping of plant responses to soil water deficit in *Arabidopsis thaliana* permitted the identification of an accession with low sensitivity to soil water deficit. *New Phytol* 169:623–635
- Gregory PJ, Hutchison DJ, Read DB et al (2003) Non-invasive imaging of roots with high resolution X-ray micro-tomography. *Plant Soil* 255:351–359
- Guo JM, Trotter CM (2004) Estimating photosynthetic light-use efficiency using the photochemical reflectance index: variations among species. *Funct Plant Biol* 31:255–265
- Haboudane D, Miller JR, Pattey E et al (2004) Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies: modeling and validation in the context of precision agriculture. *Remote Sens Environ* 90:337–352
- Hansen PM, Schjoerring JK (2003) Reflectance measurement of canopy biomass and nitrogen status in wheat crops using normalized difference vegetation indices and partial least squares regression. *Remote Sens Environ* 86:542–553
- Hargreaves CE, Gregory PJ, Bengough AG (2009) Measuring root traits in barley (*Hordeum vulgare* ssp. *vulgare* and ssp. *spontaneum*) seedlings using gel chambers, soil sacs and X-ray microtomography. *Plant Soil* 316:285–297
- Heeraman DA, Hopmans JW, Clausnitzer V (1997) Three dimensional imaging of plant roots in situ with X-ray computed tomography. *Plant Soil* 189:167–179
- Hillnhütter C, Sikora RA, Oerke E-C, Dusschoten D van (2012) Nuclear magnetic resonance: a tool for imaging below-ground damage caused by *Heterodera schachtii* and *Rhizoctonia solani* on sugar beet. *J Exp Bot* 63(1):319–327
- Hostert P, Diermayer E, Damm A, Schiefer S (2005) Spectral unmixing based on image and reference endmembers for urban change analysis. In: 24th Symposium of the European-Association-of-Remote-Sensing-Laboratories (EARSel), May 25-27, 2004, Dubrovnik. New strategies for European remote sensing, pp 645–652
- Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments. *Ecol Monogr* 54:187–211
- Iyer-Pascuzzi AS, Symonova O, Mileyko Y et al (2010) Imaging and analysis platform for automated phenotyping and trait ranking of plant root systems. *Plant Physiol* 152:1148–1157
- Jackson RD, Huete AR (1991) Interpreting vegetation indexes. *Prev Vet Med* 11:185–200
- Jahnke S, Menzel MI, van Dusschoten D et al (2009) Combined MRI-PET dissects dynamic changes in plant structures and functions. *Plant J* 59(4):634–644
- Jansen M, Gilmer F, Biskup B et al (2009) Simultaneous phenotyping of leaf growth and chlorophyll fluorescence via GROWSCREEN FLUORO allows detection of stress tolerance in *Arabidopsis thaliana* and other rosette plants. *Funct Plant Biol* 36:902–914

- Knipling EB (1970) Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation. *Remote Sens Environ* 1(3):155–159
- Kolber Z, Klimov D, Ananyev G et al (2005) Measuring photosynthetic parameters at a distance: laser induced fluorescence transient (LIFT) method for remote measurements of PSII in terrestrial vegetation. *Photosynth Res* 84:121–129
- Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu Rev Plant Biol* 55:141–172
- Kovacevic B, Roncevic S, Miladinovic D et al (2009) Early shoot and root growth dynamics as indicators for the survival of black poplar cuttings. *New Forest* 38:177–185
- Kümmerlen B, Dauwe S, Schmundt D, Schurr U (1999) Thermography to measure water relations of plant leaves Volume 3, systems and applications. In: Jähne B, Haussecker H, Geissler P (eds). *Handbook of computer vision and applications*. Academic, pp 763–781
- Malenovský Z, Mishra KB, Zemek F et al (2009) Scientific and technical challenges in remote sensing of plant canopy reflectance and fluorescence. *J Exp Bot* 60:2987–3004
- Massonnet C, Vile D, Fabre J et al (2010) Probing the reproducibility of leaf growth and molecular phenotypes: a comparison of three *Arabidopsis* accessions cultivated in ten laboratories. *Plant Physiol* 152:2142–2157
- Meininger M, Jakob PM, von Kienlin M et al (1997) Radial spectroscopic imaging. *J Magn Reson* 125(2):325–331
- Menzel MI, Oros-Peusquens A-M, Pohlmeier A et al (2007) Comparing 1H-NMR imaging and relaxation mapping of German white asparagus from five different cultivation sites. *J Plant Nutr Soil Sci* 170:24–38
- Merzlyak MN, Gitelson AA, Pogosyan SI et al (1997) Reflectance spectra of plant leaves and fruits during their development, senescence and under stress. *Russ J Plant Physiol* 44:614–622
- Merzlyak MN, Gitelson AA, Chivkunova OB, Rakitin VYU (1999) Non-destructive optical detection of pigment changes during leaf senescence and fruit ripening. *Physiol Plantarum* 106(1):135–141
- Mittler R, Blumwald E (2010) Genetic engineering for modern agriculture: challenges and perspectives. *Annu Rev Plant Biol* 61:443–462
- Moradi AB, Carminati A, Vetterlein D et al (2011) Three-dimensional visualization and quantification of water content in the rhizosphere. *New Phytol* 192:653–663
- Moya I, Camenen L, Evain S et al (2004) A new instrument for passive remote sensing 1. Measurements of sunlight-induced chlorophyll fluorescence. *Remote Sens Environ* 91:186–197
- Munns R, James RA, Sirault XRR et al (2010) New phenotyping methods for screening wheat and barley for beneficial responses to water deficit. *J Exp Bot* 61:3499–3507
- Myneni RB, Ross J, Asrar G (1989) A review on the theory of photon transport in leaf canopies. *Agr Forest Meteorol* 45:1–153
- Nagel KA, Kastenholz B, Jahnke S et al (2009) Temperature responses of roots: impact on growth, root system architecture and implications for phenotyping. *Funct Plant Biol* 36:947–959
- Nagel KA, Putz A, Gilmer et al (2012) GROWSCREEN-Rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. *Funct Plant Biol*. doi:10.1071/FP12023 39(11):891–904
- Nakazawa M, Ichikawa T, Ishikawa A et al (2003) Activation tagging, a novel tool to dissect the functions of a gene family. *Plant J* 34:741–750
- O'Malley RC, Ecker JR (2010) Linking genotype to phenotype using the *Arabidopsis* unimutant collection. *Plant J* 61:928–940
- Osmond CB, Daley PF, Badger MR, Lüttge U (1998) Chlorophyll fluorescence quenching during photosynthetic induction in leaves of *Abutilon striatum* Dicks. infected with *Abutilon* mosaic virus, observed with a field-portable imaging system. *Bot Acta* 111:390–397
- Passioura J (2010) Scaling up: the essence of effective agricultural research. *Funct Plant Biol* 37:585–591
- Pierret A, Kirby M, Moran C (2003) Simultaneous X-ray imaging of plant root growth and water uptake in thin-slab systems. *Plant Soil* 255:361–373

- Pigliucci M (2008) Ecology and evolutionary biology of *Arabidopsis*. *Arabidopsis* Book 1:e0003. doi:10.1199/tab.0003
- Purdue University (2011) 101 ways to grow *Arabidopsis*. <http://www.hort.purdue.edu/hort/facilities/greenhouse/101exp.shtml>. Accessed 1 Dec 2011
- Rascher U, Nichol CJ, Small C, Hendricks L (2007) Monitoring spatio-temporal dynamics of photosynthesis with a portable hyperspectral imaging system. *Photogramm Eng Rem Sens* 73:45–56
- Rascher U, Agati G, Alonso L et al (2009) CEFLES2: the remote sensing component to quantify photosynthetic efficiency from the leaf to the region by measuring sun-induced fluorescence in the oxygen absorption bands. *Biogeosciences* 6:1181–1198
- Rascher U, Damm A, van der Linden S et al (2010) Sensing of photosynthetic activity of crops. In: EC et al O (eds) *Precision crop protection—the challenge and use of heterogeneity*. Springer Science + Business Media BV, pp 87–99. doi:10.1007/978-90-481-9277-9_6
- Rascher U, Blossfeld S, Fiorani F et al (2011) Non-invasive approaches for phenotyping of enhanced performance traits in bean. *Funct Plant Biol* 38:968–983
- Reboud X, Le Corre V, Scarcelli N et al (2004) Natural variation among accessions of *Arabidopsis thaliana*: beyond the flowering date, what morphological traits are relevant to study adaptation? In: Cronk QCB, Whitton J, Ree RH, Taylor IEP (eds) *Plant adaptation: molecular genetics and ecology*. Natl Research Council Canada, Ottawa, pp 135–142
- Richards RA (2000) Selectable traits to increase crop photosynthesis and yield of grain crops. *J Exp Bot* 51:447–458
- Rogers HH, Bottomley PA (1987) *In situ* magnetic resonance imaging of roots: influence of soil type, ferromagnetic particle content, and soil water. *Agron J* 79:957–965
- Rokitta M, Peuke AD, Zimmermann U, Haase A (1999) Dynamic studies of phloem and xylem flow in fully differentiated plants by fast nuclear-magnetic-resonance microimaging. *Protoplasma* 209:126–131
- Rollin EM, Milton EJ (1998) Processing of high spectral resolution reflectance data for the retrieval of canopy water content information. *Remote Sens Environ* 65(1):86–92
- Römer C, Wahabzada M, Ballvora A et al (2012) Early drought stress detection in cereals: simplex volume maximization for hyperspectral image analysis. *Funct Plant Biol* 39:878–890
- Schilling M, Pfeifer AC, Bohl S, Klingmüller U (2008) Standardizing experimental protocols. *Curr Opin Biotech* 19:354–359
- Simpson AJ, McNally DJ, Simpson MJ (2011) NMR spectroscopy in environmental research: from molecular interactions to global processes. *Prog Nucl Magn Reson Spectrosc* 58:97–175
- Skirycz A, Vandenbroucke K, Clauw P et al (2011) Survival and growth of *Arabidopsis* plants given limited water are not equal. *Nat Biotechnol* 29:212–214
- Stylinski CS, Gamon JG, Oechel WO (2002) Seasonal patterns of reflectance indices, carotenoid pigments and photosynthesis of evergreen chaparral species. *Oecologia* 131(3):366–374
- Sultan SE (2000) Phenotypic plasticity for plant development, function and life history. *Trends Plant Sci* 5:537–542
- Turner DP, Cohen WB, Kennedy RE et al (1999) Relationships between leaf area index and landsat TM spectral vegetation indices across three temperate zone sites. *Remote Sens Environ* 70:52–68
- Ustin S, Gamon JA (2010) Remote sensing of plant functional types. *New Phytol* 186:795–816
- As H van (2007) Intact plant MRI for the study of cell water relations, membrane permeability, cell-to-cell and long distance water transport. *J Exp Bot* 58:743–756
- As H van, Scheenen T, Vergeldt FJ (2009) MRI of intact plants. *Photosynth Res* 102:213–222
- Verrelst J, Schaepman ME, Koetz B, Kneubühler M (2008) Angular sensitivity analysis of vegetation indices derived from CHRIS/PROBA data. *Remote Sens Environ* 112:2341–2353
- Walter A, Schurr U (2005) Dynamics of leaf and root growth: endogenous control versus environmental impact. *Ann Bot* 95:891–900
- Walter A, Rascher U, Osmond CB (2004) Transition in photosynthetic parameters of midvein and interveinal regions of leaves and their importance during leaf growth and development. *Plant Biol* 6:184–191

- Walter A, Scharr H, Gilmer F et al (2007) Dynamics of seedling growth acclimation towards altered light conditions can be quantified via GROWSCREEN: a setup and procedure designed for rapid optical phenotyping of different plant species. *New Phytol* 174:447–455
- Walter A, Silk WK, Schurr U (2009) Environmental effects on spatial and temporal patterns of leaf and root growth. *Annu Rev Plant Biol* 60:279–304
- Weigel D, Glazebrook J (2002) *Arabidopsis: a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor
- Zadoks JC, Chang TT, Konzak CF (1974) A decimal code for the growth stages of cereals. *Weed Res* 14:415–421 and *Eucarpia Bull* 7:49–52
- Zhu J, Ingram PA, Benfey PN, Elich T (2011) From lab to field, new approaches to phenotyping root system architecture. *Curr Opin Plant Biol* 14:310–317

Chapter 9

Association Mapping of Genetic Resources: Achievements and Future Perspectives

Sivakumar Sukumaran and Jianming Yu

Contents

9.1	Introduction	208
9.1.1	Population Structure and Association Mapping Methods	209
9.1.2	Nested Association Mapping (NAM)	212
9.1.3	Software for Association Mapping	213
9.1.4	Computational Speed	214
9.2	Achievements	215
9.2.1	Association Mapping in Plants	215
9.2.2	GWAS in Plants	216
9.2.3	Arabidopsiss	216
9.2.4	Maize	218
9.2.5	Rice	221
9.2.6	Community Resources in Wheat, Barley, Soybean, and Sorghum	222
9.3	Challenges and Opportunities	226
9.3.1	Missing Heritability	228
9.3.2	New Gene Identification	228
9.3.3	Genotyping-by-Sequencing (GBS)	229
9.3.4	Rare Alleles	229
9.3.5	Genic and Nongenetic Contribution	230
	References	230

Abstract Association mapping studies in plants contribute to not only detecting the genetic basis of variation in physiological, developmental, and morphological traits (e.g., flowering time, plant height, grain quality, and nutrient content) but also bringing together researchers to assemble core collections and develop genetic platforms for genotyping, phenotyping, analysis, and interpretation. The establishment of the unified mixed model greatly facilitated association mapping studies in plants and further methodology work in general. Association mapping is well positioned to exploit the advances in next generation genomic technologies and high-throughput

J. Yu (✉)

Department of Agronomy, Iowa State University, Ames, Iowa, USA
e-mail: jmyu@iastate.edu

S. Sukumaran

International Maize and Wheat Improvement Center (CIMMYT),
Apdo. Postal 6-641, 06600 Mexico DF, Mexico
e-mail: s.sukumaran@cgiar.org

phenotyping. Genome-wide association studies (GWAS) are expected to increase dramatically once genome sequences of all major crop species are obtained. Moving forward, researchers in plant genetics and related disciplines need to develop improved genetic designs and computational tools to address several challenges such as missing heritability, new gene identification, genotyping-by-sequencing, and rare alleles. In this chapter, we describe major progress in understanding population structure, advancements in design and implementation of association mapping, and summarize examples of association mapping in maize, rice, *Arabidopsis*, wheat, barley, soybean, and sorghum. Finally, major opportunities with potential implications in plant genetics are discussed.

Keywords Association mapping · Genome-wide association studies (GWAS) · Genetic resource · Missing heritability · Rare allele · Population structure · Nested association mapping (NAM) · Genotyping-by-sequencing (GBS) · Genetic diversity · Linkage disequilibriums

9.1 Introduction

The priority of the plant genetics and breeding research community is to increase yield and stability of major crops to ensure food security for the fast growing population. To meet the challenge, plant breeding methods, genetic designs, genomics, and biotechnologies need to be integrated to modify the adaptive, agronomic, and economic characteristics. Two connected components of this endeavor are gene identification and complex trait dissection. While gene identification focuses on individual genes, complex trait dissection emphasizes genetic contribution and modes of action from many loci that result in phenotypic variation. Linkage mapping and association mapping are the most commonly used approaches in dissecting complex traits and identifying genes underlying trait variation in plants, animals, and human (Risch and Merikangas 1996).

Association mapping provides a great platform to exploit genomic technologies and plant germplasm resources simultaneously (Zhu et al. 2008). Compared with the traditional bi-parental linkage analysis, association mapping offers several advantages. Association mapping populations are typically assembled with diverse lines from breeding programs or sampled accessions from germplasm banks. As a result, researchers can initiate genotyping and phenotyping activities with this approach while developing complementary linkage mapping populations. Because association mapping utilizes the historic recombinations present in the panel, a higher mapping resolution is expected. The approach is also fast, and it can accommodate a large number of accessions and analyze a higher number of alleles (Zhu et al. 2008; Myles et al. 2009).

Based on the scale of the research, association mapping can be classified into either targeted candidate-genes studies or large-scale genome-wide association studies

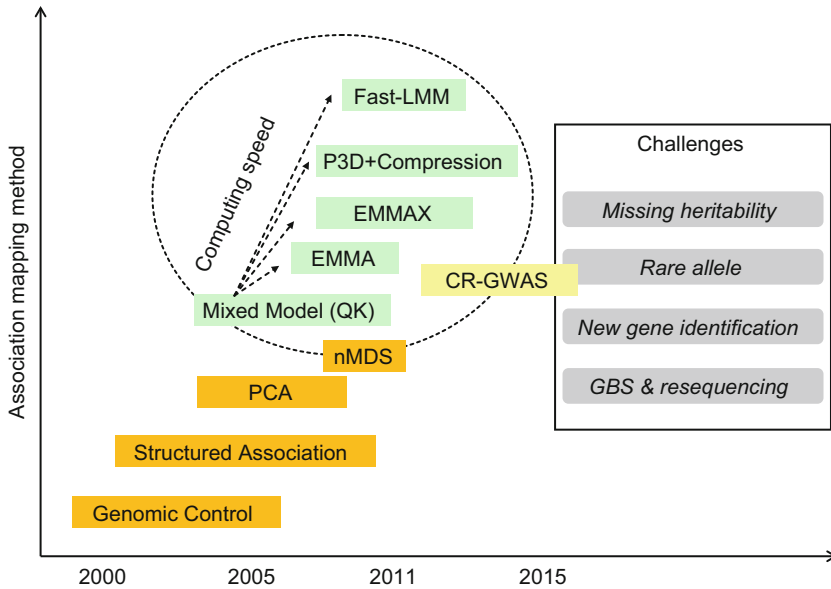


Fig. 9.1 Achievements in association mapping methods and future challenges

(GWAS) (Zhu et al. 2008). A combined approach also can be taken in which a specific genomic region is subjected to a high-resolution scan with a large number of markers. The candidate gene approach focuses on specific traits and targeted genes with known biochemical pathways and other information. In GWAS with high genome coverage of SNP markers, it is possible to identify genes with previously unknown functions. In addition, another integrated approach, Nested Association Mapping (NAM), has been implemented in plants to combine the advantages of linkage analysis and association mapping (Yu et al. 2008). Although earlier association mapping studies focused mainly on candidate genes (Thornsberry et al. 2001) with limited sample sizes and marker numbers, recent advances in analysis methods and genomic technologies have enabled GWAS in many plant species (Fig. 9.1). We encourage interested readers to refer to previously published reviews (Flint-Garcia et al. 2003; Zhu et al. 2008) for specific details of linkage disequilibrium and association mapping. In this chapter, we will focus on major achievements: understanding of the population structure, research strategies, and examples and progress in major crops. We then outline the main challenges that have broad implications.

9.1.1 Population Structure and Association Mapping Methods

Because an association mapping panel is often an assembled population, rather than a random mated or a designed population, the presence of population structure (*i.e.*,

unequal genetic relationship among groups of individuals) can lead to false positive discoveries if this structure is not adequately accounted for during marker-trait association analysis. In general, population structure can arise due to differences in geographical origins, local adaptations, or breeding history of the lines in the panel (Yu et al. 2006). One important note for the term “population structure” is that when it is used alone, we often refer to the unequal genetic relationship among individuals of the association mapping panel. But when this term is used together with “relative kinship” or “familial relationship”, we often refer to the general grouping pattern of the individuals that are captured by the STRUCTURE analysis.

As is often true in many crops, multiple levels of relatedness frequently exist in association mapping populations as a result of assembling lines and accessions from different geographical regions or breeding programs (*i.e.*, major grouping patterns), as well as different levels of relationship among individuals within individual breeding programs (*i.e.*, pairwise relationship). In such cases, some markers may appear to be significantly associated with the trait of interest with a simple test, but their frequency distributions are in fact correlated with the population structure. If there is an average trait difference among groups of individuals, this frequency distribution would then lead to the significance of these markers. Therefore, understanding the structure of the population, developing methods to accurately infer the structure, and conducting association analysis with appropriate models are critical to reduce or eliminate these spurious associations or false positives.

Association mapping samples generally fall into five categories based on population structure and familial relatedness. Population structure is associated with local adaptation or diversifying selection, and the familial relationship is associated with the recent coancestry. The five categories are (1) ideal samples with subtle population structure and familial relatedness, (2) samples with familial relationship, (3) samples with population structure, (4) samples with both population structure and familial relationships, and (5) samples with severe population structure and familial relationships (Zhu and Yu 2009). It is possible to quantify population structure using neutral markers and then account for the structure statistically in identifying marker-trait associations. Several methods have been used to control for population structure in association mapping (Fig. 9.1). These include genomic control (Devlin and Roeder 1999), structured association (Pritchard et al. 2000b), principal component analysis (PCA) (Patterson et al. 2006; Price et al. 2006), unified mixed model (Yu et al. 2006), non-metric multidimensional scaling (nMDS) (Zhu and Yu 2009), and techniques to increase computational speed and power of the mixed model (Zhang et al. 2010b).

Recent methods mainly use the mixed linear model (MLM) to account for population structure and familial relatedness. The unified mixed model (Yu et al. 2006) approach for association mapping considers both population structure and pairwise relatedness to account for genetic relationship. With this general framework, population structure (Q matrix) estimated using STRUCTURE software (Pritchard et al. 2000a) is fitted as a fixed effect, and kinship (K matrix) estimated using SPAGeDi (Hardy and Vekemans 2002) is used to define the variance-covariance structure of the random effects among individuals for association mapping. This K estimates the identity by descent (IBD) by adjusting the probability of identity by state (IBS)

different numbers of PCAs are needed to have adequate, but not excessive, control for population structure.

With the unified mixed model framework, newer and faster algorithms have been developed to speed up GWAS analysis, particularly when hundreds of thousands of SNPs are tested. Efficient mixed-model association (EMMA) corrects sample structures by accounting for pairwise relatedness between individuals and uses enough markers by modeling phenotype distribution. This method is related to a method developed to simulate a null distribution of variance component test statistics (Crainiceanu and Ruppert 2004). EMMA increases the computational speed and efficiency of mixed model analysis by enabling statistical tests with single-dimensional optimization. The method also avoids the redundant, computationally expensive matrix operations at iteration and allows converge to the global optimum of the likelihood in variance component estimation with high confidence. This capability was demonstrated in *in silico* whole-genome association mapping of mouse, *Arabidopsis*, and maize datasets. Results from the EMMA method are consistent with published results in reducing false positives and are faster than the previous methods while performing near global optimization (Kang et al. 2008, 2010). Compressed MLM is another approach that clusters individuals into groups based on kinship estimates, thereby reducing the effective sample size for computation to improve speed in subpopulation determination. This method is an extension of the pedigree-based sire model (Henderson 1975) with modifications (Zhang et al. 2010b).

9.1.2 Nested Association Mapping (NAM)

Ideally, an association mapping population can be genotyped once but phenotyped repeatedly for the same sets of traits and new sets of traits; thus, it is advantageous to have a population that can be used by the research community for different purposes. With this in mind, the maize community has developed a nested association mapping (NAM) population to integrate the advantages of linkage analysis and association mapping, with the ultimate goal of dissecting complex traits in maize (Yu et al. 2008). The aims of developing NAM population were to (1) capture maize genetic diversity, (2) exploit the historical recombinations in maize, (3) use a genetic design that can take advantage of next-generation sequencing technologies, (4) generate materials for evaluation of agronomic traits in the field locations of the temperate regions, (5) develop a population with enough power to detect QTLs and resolve QTLs to the gene level, and (6) provide a community resource that will enable a wide range of community efforts and databases for researchers. A publicly available resource of immortal lines with 26 founders represents the global diversity of maize. A set of 25 diverse inbred lines were crossed to common reference maize inbred line B73 to create 25 populations of 200 RILs each, for a total of 5,000 distinct genotypes. The 5,000 genotypes are called NAM recombinant inbred lines (NAM RILs).

In essence, NAM exploits a multiple RIL population derived from crosses between a common founder line and a set of diverse founders. The strategy of NAM is to genotype common-parent-specific (CPS) markers on the founders and progenies

while sequencing the founders completely or densely genotyping the founders with high-density markers. With CPS markers serving as a bridge, the genetic information obtained from genotyping the founders with high-density markers can be projected from the founders to the progenies. Projecting genetic information from the parents to the progenies also reduces genotyping costs. Notice that the concept of NAM involves the development of a population. Instead of assembling existing lines to form a population, NAM selects a diverse set of founders that are representative of the main breeding pools of the target species. Importantly, as compared with the conventional association mapping approach, NAM is characterized by higher statistical power, better mapping resolution, lower sensitivity to genetic heterogeneity, and a lower requirement of SNP markers in the progenies (Yu et al. 2008). In maize, a number of studies have been conducted on the genetic map properties of the NAM populations (McMullen et al. 2009), flowering time (Buckler et al. 2009), leaf architecture (Tian et al. 2011), and disease resistance (Kump et al. 2011; Poland et al. 2010). Some of these studies will be further reviewed in section 9.2.4.

9.1.3 Software for Association Mapping

In this section, we will mention new algorithms implemented in a set of software packages. Some detailed information about these algorithms are further explained in the next section. The most commonly used and frequently updated software for association mapping is TASSEL (Trait Analysis by Association, Evolution and Linkage), which is written in Java and can be used in virtually any operating system (Bradbury et al. 2007). TASSEL implements GLM, MLM, compressed MLM, and P3D approaches for marker-trait association analysis. Other notable functions include evolutionary analysis, computation of LD, imputation of missing data, and data visualization. The program allows calculating and visualizing LD graphically.

Structured association (Pritchard et al. 2000b) as well as the unified mixed model (Yu et al. 2006) were first implemented in TASSEL to reduce the risk of false positives. The $Q + K$ method was implemented in TASSEL as a MLM function. TASSEL earlier employed an EM (expectation-maximization) algorithm for MLM analysis. To increase computing speed and analyze larger datasets, the EMMA algorithm was incorporated into TASSEL. As indicated earlier, compressed MLM was added recently to increase computational speed and this procedure can also be optimized to increase the power.

For even larger datasets, a newer method estimates the population parameters prior to estimating the parameters for test markers. Termed as population parameters previously determined (P3D) (Zhang et al. 2010b), this method is available in the newer version of TASSEL, and compressed estimation of variance components are available in software EMMA eXpedited (EMMAX) (Kang et al. 2010). EMMAX, a publicly available software, implements a variance-component approach for GWAS. EMMAX is built on EMMA's previous approach.

ASREML is a complete package with different modules for mixed model analysis (Gilmour et al. 2002). SAS and R software are generic tools that can be used for association mapping. GAPIT (Genome Association and Prediction Integrated Tool) is a new tool in the R package that can perform genome-wide association study (Lipka et al. 2012). It integrates the unified mixed model, EMMA, P3D, and compressed MLM with genomic prediction. This software handles large genotypic datasets by subdividing them into multiple files, but the memory requirement remains the same. Genomic predictions are done using a method developed by Zhou et al. (2011). GAPIT can conduct hierarchical clustering and kinship matrices based on user input and linkage information. The results are produced in the form of Q-Q plot, Manhattan plot, PCA, and association tables.

Genome-wide efficient mixed model association (GEMMA) is a method n (sample size) times faster than the EMMA method. It is not an approximation method similar to genome-wide rapid association using mixed model and regression (GRAMMER) but an exact test method. This requires complete or imputed SNP data and for each SNP tested it replaces the additional eigen-decomposition step in EMMA with one-matrix vector multiplication. After that, each iteration of the optimization requires inexpensive operations. GEMMA was built on EMMA framework to facilitate genome-wide application (Zhou and Stephens 2012).

Multi-locus mixed model (MLMM) is a method which can outperform the existing methods in analyzing GWAS data in power and false discovery rate. The existing methods to account for population structure are good when small number of loci control the traits. But for complex traits controlled by several large-effect loci MLMM is efficient. The principle of this approach is similar to the use of cofactors in the statistical models of QTL mapping in composite interval mapping multiple interval mapping. Likewise including multiple cofactors on a genome-wide scale can increase the power and reduce false discovery rates in GWAS (Segura et al. 2012).

9.1.4 Computational Speed

The unified mixed model method originally developed by Yu et al. (2006) is widely used to correct for the effects of genetic relatedness in association mapping studies; however, in analyzing genome-wide datasets, solving the mixed model requires a huge amount of computing power. The computing time for solving an MLM increases with the cube of the number of individuals. One approach to reducing computing time is compressed MLM (Zhang et al. 2010b), which decreases the effective sample size of such datasets by clustering individuals into groups. The rationale behind this method has its roots in the sire-model approach (Quaas and Pollak 1981).

As a complementary approach to compressed MLM, the population parameters previously determined (P3D) approach reduces computing time by skipping the iteration process in each individual marker test. In the first step, a base MLM without fitting any marker effect is solved for the variance components. In the second step, an individual marker test with MLM simply uses variance components from the first step

without solving the specific mixed model again (Zhang et al. 2010b). This practice has been used in previous mixed model analyses to save computing time (Yu et al. 2005), but the need to reduce the computational burden of MLM is much higher in GWAS. Compressed MLM and P3D, when implemented jointly, significantly reduce computing time and maintain statistical power (Zhang et al. 2010b). These methods are implemented in the software program TASSEL (Bradbury et al. 2007). A different residual analysis approach was also proposed to conduct fast genome-wide pedigree-based association analysis (Aulchenko et al. 2007).

A variance component approach implemented in a publicly available software package, EMMAX (Kang et al. 2010), reduces computing time for analyzing large GWAS datasets. First, a pairwise relatedness matrix is computed from high-density markers and used to represent the sample structure. Secondly, the contribution of the sample structure to the phenotype using a variance component model is estimated, resulting in an estimated covariance matrix of phenotypes that models the effect of genetic relatedness on the phenotypes. Thirdly, a generalized least square (GLS) F -test (Kariya and Kurata 2004) is applied to each marker to detect associations accounting for the sample structure using the covariance matrix. A study on the welcome trust consortium data (Browning and Browning 2008) found that EMMAX outperforms both PCA (Price et al. 2006) and genomic control (Devlin et al. 2001).

FaST-LMM, a factored spectrally transformed linear mixed model, was recently proposed to further address the computational issues of MLM (Lippert et al. 2011). FaST-LMM is an algorithm for genome-wide association studies that scales linearly with sample size in both runtime and memory use. With data from 15,000 individuals, FaST-LMM ran an order of magnitude faster than current algorithms; whereas data for 120,000 individuals were analyzed with FaST-LMM in few hours, current algorithms failed. The LMM corrects for confounding by measuring genetic similarity using methods of identity by descent and a realized relationship matrix (RRM), estimated by using a small sample of markers. FaST-LMM can produce results similar to the LMM by reformulating the LMMs with two conditions: (1) the number of SNPs used to estimate genetic similarity is less than the number of individuals in the dataset, and (2) the RRM is used to determine these similarities. This method requires a single spectral decomposition but does not assume variance parameters are same across the SNPs.

9.2 Achievements

9.2.1 Association Mapping in Plants

Association mapping in plants provides a powerful, complementary approach to existing QTL mapping and cloning with bi-parental populations, mutational dissection, and transgenic approaches. It has been widely adopted in almost all major crop species for gene identification and QTL validation, and to better understand the genetic basis of complex traits (Zhu et al. 2008). Association mapping also has led

to the development of common community resources in important crop species, including barley, maize, rice, sorghum, soybean, and wheat. Linkage disequilibrium estimates among a reasonably large and diverse set of accessions within a species typically provide basic knowledge about the potential resolution of association mapping and the marker density requirement within each set (e.g., elite materials vs. landraces). Following these LD studies, population structure analysis of the assembled association mapping is examined in details. The resulting information is then incorporated into either candidate-gene or genome-wide association analysis.

One of the major benefits of association mapping is the diversity captured across many different traits. Unlike specific bi-parental populations in which certain trait differences exist, most of the assembled association mapping panels can be used to study a host of traits so that questions from different angles can be studied, including basic biology, plant architecture, development, agronomic performance, adaptive characteristics, and nutritional value (Atwell et al. 2010; Flint-Garcia et al. 2005).

9.2.2 GWAS in Plants

GWAS is a routinely adopted approach in human disease studies and has been carried out with success also in plants. To dissect complex traits through whole-genome association mapping, diverse germplasm panels have been established in *Arabidopsis* (Nordborg et al. 2005), barley (Caldwell et al. 2006), maize (Yu and Buckler 2006), rice (Huang et al. 2010, 2011, 2012,) sorghum (Casa et al. 2008), soybean (Lu et al. 2011), wheat (Brescghello and Sorrells b), and durum wheat (Maccaferri et al. 2005, 2006) (Fig. 9.2). The number of GWAS report has been limited in plants compared with that in human diseases. In next section, we describe some of the experiments that have been successful in detecting marker-trait associations following the GWAS strategy in plants (Table 9.1).

9.2.3 *Arabidopsis*

Arabidopsis thaliana is a natural organism that exists in a wide range of habitats. *Arabidopsis* HapMap is a resource to study the evolutionary as well as functional genetics of natural populations to resolve complex trait variation due to individual genes and, in some cases, the individual nucleotides (Clark et al. 2007). Based on the genome analysis, LD decays rapidly in this species, within 50 kb (Platt et al. 2010). As a model species with a small genome, *Arabidopsis* has been a frontrunner in association studies. The 1001 Genomes project is sequencing 1001 geographically diverse *Arabidopsis* strains (Weigel and Mott 2009). In the first phase of the project, 80 strains from eight regions of the native species range were sequenced and analyzed (Cao et al. 2011). Another report claims that 471 genomes have been sequenced (unpublished data). Two recent GWAS studies in *Arabidopsis* are reviewed below.

Table 9.1 Examples of association mapping study in different crops

Species	Trait/objective	Population	Marker	Reference
Arabidopsis	107 phenotypes	191 accessions	250,000 SNPs	Atwell et al. 2010
Arabidopsis	Flowering time	184 accessions + 4,366 RILs	216,509 SNPs	Branchi et al. 2010
Zea mays	Leaf angle, leaf length, and width	Maize NAM	1.6 million SNPs	Tian et al. 2011
Zea mays	Southern leaf blight disease	Maize NAM	1.6 million SNPs	Kump et al. 2011
Zea mays	Provitamin A	288 lines	Candidate genes	Harjes et al. 2008
Zea mays	Provitamin A	681 maize germplasm	Candidate genes	Yan et al. 2010
Oryza sativa	14 agronomic traits	517 landraces	3.6 million SNPs	Huang et al. 2010
Oryza sativa	Flowering time and 10 grain-related traits	950 worldwide accessions	4.1 million SNPs	Huang et al. 2011
Triticum aestivum	To understand genetic diversity, population structure, and linkage disequilibrium	205 elite breeding lines	245 SSRs	Zhang et al. 2010a
Triticum durum	Drought-adaptive traits and grain yield	189 elite durum	186 SSRs	Maccaferri et al. 2011
Triticum durum	Resistance to stem rust	183 elite durum	323 SSRs	Letta et al. 2013
Hordeum vulgare	15 morphological traits	500 cultivars + DH populations	538 DaRT markers	Cockram et al. 2010
Hordeum vulgare	Domestication traits	190 cultivars	1536 SNPs 2463 SNPs	Ramsay et al. 2011

The feasibility of GWAS in plants was demonstrated by studying a sample *Arabidopsis thaliana* global population. The genotyped sample consisted of 95 accessions for which a number of phenotypes were available (Zhao et al. 2007), plus a set of 96 accessions for which flowering traits were available (Atwell et al. 2010). The genotyping chip containing 250,000 SNPs was used to genotype the accessions; thereby, the SNP density was one SNP per 500 bp, which is comparable to studies in humans. The phenotypes studied were related to flowering, plant defense, mineral concentrations, and developmental traits. This association sample has a highly complex population structure. The mixed model approach performed well in controlling the population structure compared with other methods commonly used in human genetics. Even though the degree of confounding was different among phenotypes, association analysis effectively identified single genes with known functional polymorphism. Notwithstanding the small sample size, particularly when compared with the human GWAS, the gene identification was possible due to the genetic architecture of the trait. These studies are replicable under controlled conditions that eliminate environmental noise (Atwell et al. 2010).

Another comprehensive study of linkage and association mapping of flowering time was conducted under field and greenhouse conditions (Brachi et al. 2010). The experiment involved phenotyping 20,000 plants over two winters under field conditions. A set of 184 natural accessions from around the world was genotyped with 216,509 SNPs, and 4,366 RILs derived from 13 independent crosses were also examined. More than 60 QTLs with small to medium effects were identified. The highlight of this study was that linkage mapping, which has a higher power to distinguish true positives from false positives than association mapping, should be integrated with GWAS. A second important finding was that the major genes governing flowering time in greenhouse conditions were not associated with flowering time in field conditions. Instead, a number of genes involved in the regulation of the plant circadian clock were associated (Brachi et al. 2010).

9.2.4 Maize

The genetic diversity between two different maize inbred lines is roughly equivalent to the diversity between a man and a chimp (Buckler et al. 2006). The maize genome contains about 50,000 genes, and most of the genome comprises repetitive and transposable elements (Schnable et al. 2009). The huge genetic diversity of maize make high-resolution mapping possible, but also requires large numbers of SNPs and systematic analysis (Yu and Buckler 2006). Several maize association mapping populations have been assembled (Remington et al. 2001; Liu et al. 2003; Palaisa et al. 2003; Flint-Garcia et al. 2005; Camus-Kulandaivelu et al. 2006; Yu et al. 2008). In addition to studies in which diverse germplasm was exploited first, association mapping has also been effectively deployed to confirm the putative role of the *Vgt1* locus in controlling flowering time in maize after the QTL cloning process (Salvi et al. 2007).

The Maize HapMap project is an excellent resource for plant geneticists to conduct association mapping. The first HapMap in maize (Gore et al. 2009) identified several million polymorphisms (1.4 million SNPs and 200,000 indels) among 27 diverse maize inbred lines and showed that the maize genome is characterized by highly divergent haplotypes. The second maize HapMap (HapMapV2) resulted in the identification of high-quality genotypic data of 50 million SNPs and small indels (Chia et al. 2012). These efforts provide the foundation for dissecting the complex trait variation in maize by uniting research efforts around the world.

To determine whether standing variation in the regulatory genes in maize contributes to variation in *Balsas teosinte*, association mapping was conducted on 584 *Balsas teosinte* individuals. Forty-eight markers from 9 candidate regulatory genes were tested against 13 traits for plant and inflorescence architecture. Ten associations involving five candidate genes were significantly identified after correcting for multiple testing. The maize homolog of *FLORICAULA* of *Antirrhinum zfl2* was associated with plant height. The maize homolog of *APETALA1* of *Arabidopsis zap1* was associated with inflorescence branching. Five SNPs in the maize domestication gene, *teosinte branched1*, were significantly associated with either plant or inflorescence architecture (Weber et al. 2007).

To address vitamin A deficiency, the leading cause of blindness, disease, and death from severe infections in children. Breeding for increased β -carotene (β C) levels in cereal grains (biofortification) is a realistic approach because β -carotene is a precursor of vitamin A. In the first study, association mapping coupled with linkage analysis, expression analysis, and mutagenesis identified variation in the *lycopene epsilon cyclase* (*lcyE*) gene that accounts for 58 % of the variation in the α - vs. β -carotene branches of the carotenoid pathway and a threefold difference in provitamin A compounds. The *lcyE* gene was significantly associated with the branching and carotenoid content (Harjes et al. 2008).

In the second study, three association mapping populations were used. SSR and SNP markers were used to estimate the population structure and kinship matrices. GWAS identified a rare variation in the *crtRBI* gene in maize, which increases the β -carotene concentration and conversion in maize kernels (Yan et al. 2010). Results from these studies will facilitate breeding for increased β -carotene levels in cereal grains, thereby addressing the dietary vitamin-A deficiency in the developing world. This is a good example of cross-validating the QTLs using a combination of association and linkage mapping strategies.

The NAM population has been used to elucidate the genetic basis of resistance to southern leaf blight (SLB) disease. SLB resistance was measured on a nine-point scale in three environments. The SLB index values varied among the founder lines, with B73 being the least resistant. The heritability of the SLB index was 87 %, indicating potential for accurate mapping. Joint-linkage analysis identified 32 QTLs with small additive effects on SLB resistance. Genome-wide association tests of Maize HapMap were conducted by imputing the founder SNPs onto the NAM RILs; SNPs within and outside the QTLs were found to be associated with the variation for SLB resistance. Limited LD was observed around some SNPs, which indicates that NAM population is good for high-resolution mapping. But half of the QTLs

detected by the bi-parental mating studies were not detected in NAM, due to either low frequency or absence of the alleles (Kump et al. 2011).

To gain insight into the genetic architecture of quantitative resistance to plant pathogens, the NAM RILs were evaluated for resistance to northern leaf blight (NLB). Using 1.6 million SNPs, multiple candidate genes related to plant defense were identified. Twenty-nine QTLs were identified, most of which had multiple alleles. The study concluded that the large amount of variation present in the phenotype could be attributed to a number of loci with small effects (Poland et al. 2010).

Over the years, maize yield in the United States has increased partly because of reduced response to high planting density and efficient light capture, which has been possible because breeders changed the plant architecture by selecting for small leaf angle and leaf size. One study by Tian et al. focused on the genetic basis of the factors responsible for increased yield in corn. They considered the genetic basis of leaf architecture traits in maize and identified key genes through GWAS on the NAM population (Tian et al. 2011). This study demonstrated that the genetic architecture of the leaf traits (upper leaf angle, leaf length, and width) are dominated by QTLs with small effects, little epistasis, and environmental interaction or pleiotropy. The study showed that the variation in the liguleless genes has contributed to more upright leaves. For these three leaf traits, 30-36 QTLs were identified.

Flowering time is one of the traits most thoroughly studied by the plant community. It is a complex trait that controls the plant's adaptation to the local conditions. In maize, diversity-based dissection of flowering time is problematic due to tight linkage and population structure. Association mapping was used to validate the role of *Vgt1* a major QTL for flowering time originally detected and fine-mapped in biparental populations (Salvi et al. 2007). Buckler and coworkers evaluated one million plants of the NAM population for flowering time over eight environments and were able to confirm the role of *Vgt1* while identifying numerous QTLs with small effects (Buckler et al. 2009). This study evaluated 5,000 lines plus 500 checks in four environments over two years for flowering time. Days to silking (DS) and days to anthesis (DA) were scored, and anthesis silking interval (ASI) was calculated. The QTLs were mapped on the 25 families separately using composite interval mapping (CIM) and jointly by joint inclusive composite interval mapping (JICIM). JICIM identified twice as many significant QTLs as the individual family analysis. No single QTL with large effects was detected in the study. The NAM founders showed allelic differences for allelic effects. This study showed that for an adaptive trait like flowering time, the genetic architecture is controlled by small additive genes with small genetic effects and environmental interactions.

To evaluate the hypothesis that the genes controlling multiple disease resistance (MDR) is present in maize, a mixed model approach for structured association was extended to multivariate analysis. The analysis of a panel of 253 maize inbred lines identified high positive genetic correlations between resistances to southern leaf blight (SLB), northern leaf blight (NLB), and gray leaf spot (GLS). A glutathione *S*-transferase gene (*GST*) was conferring resistance to the three diseases. These successful examples in maize will encourage GWAS studies in other crops (Wisser et al. 2011).

9.2.5 Rice

Rice is a staple food for half of the world's population, and rice varieties are adapted to varied climatic regions around the globe. Rice is a highly self-fertilizing species with a high-quality reference genome (Sequencing Project International Rice 2005) and established phenotyping resources. The genome of domesticated rice contains information that could explain a large amount of the morphological, physiological, and ecological variation present in most of the cultivars throughout the world (McNally et al. 2009).

Seed shattering is a major trait in the domestication of crop plants. Konishi et al. (2006) studied seed shattering through haplotype analysis and association analysis in various rice collections. The study revealed that an SNP highly associated in *japonica* subspecies was the target for artificial selection. QTL analysis revealed that the loss of seed shattering might have occurred independently in *japonica* and *indica* varieties. A QTL for seed shattering in chromosome 1 (*qSH1*) explained 68.6 % of the total phenotypic variation in the population. Fine-mapping the *qSH1* gene with 10,388 plants identified a 612 bp region with one SNP that is responsible for the phenotype. The complementation tests proved that the *qSH1* gene was the homolog of the *RLP* gene in *Arabidopsis*. The researchers also verified the SNP using association analysis of the rice core collections, which indicated that it was highly associated with the degree of seed shattering among the temperate *japonica* rice cultivars. This SNP was a target of artificial selection in rice domestication (Konishi et al. 2006).

Apart from seed shattering, rice domestication is also associated with improvement in grain size, grain number, panicle size, grain quality, plant architecture, and flowering time, but the primary objective was grain yield. Through fine-mapping, complementation testing, expression analysis, and haplotype testing Shomura et al. (2008) found that a deletion in *qSW5* (a QTL for seed width on chromosome 5) significantly increased sink size, grain weight, and ultimately grain yield (Shomura et al. 2008).

To access the variation present within and between rice cultivars and landraces, the International Rice Functional Genomics Consortium (IRFGC) initiated an SNP discovery project (McNally et al. 2006). With this project, rice was the first crop plant for which a high-quality reference genome sequence from a single variety was produced. Through whole-genome comparisons of the 21 rice genomes including cultivars, landraces, and breeding materials (publically available in www.oryzasnp.org), 160,000 non-redundant SNPs were identified, thus providing the foundation for high-resolution genotyping of thousands of varieties (Huang et al. 2010).

Rice domestication was a complex process. The deep genetic divergence between the two main varietal groups (*indica* and *japonica*) suggests domestication of rice from two distinct wild populations. GWAS was performed in rice to understand the genome-wide patterns of polymorphism, to characterize population structure, and to infer the introgression history of domesticated Asian rice. The analysis showed that a key gene, *SD1* (*OsGA20* oxidase), determines plant height and was the target of the Green Revolution (Zhao et al. 2010).

GWAS was performed on 517 landraces of rice with 3.6 million SNPs to understand the genetic basis of diverse varieties in rice. A Rice HapMap was created and GWAS was performed for 14 agronomic traits. The LD decay of indica and japonica were between 123 and 167 kb. The simple as well as the compressed MLM models were used to identify the association signals. On average, the loci identified through GWAS explained ~36% of the phenotypic variance. The highly significant associations of six loci were close to the previously identified genes. The researchers reported that an approach which integrates the second genome sequencing and GWAS could be used as a powerful complementary strategy to traditional linkage mapping in dissecting complex traits (Huang et al. 2010).

Genome-wide association mapping revealed a rich architecture of complex traits in rice. Numerous common variants influencing physiological, developmental, and morphological traits were identified by a genome-wide association study based on genotyping 44,100 SNP variants across 413 diverse accessions of *O. sativa* collected from 82 countries that were systematically phenotyped for 34 traits. Significant heterogeneity was observed in the genetic architecture associated with subpopulation structure and response to environments. This study was an open-source translational research platform for genome-wide association studies in rice that directly linked molecular variation in genes and metabolic pathways with the germplasm resources needed to accelerate varietal development and crop improvement (Zhao et al. 2011).

Another GWAS in rice examined a diverse sample of 950 worldwide varieties that included *indica* and *japonica* subspecies. The researchers identified 32 new loci associated with flowering time and grain-related traits using the compressed MLM approach. The study reveals that an integrated approach following sequencing-based GWAS and functional genome annotation has the potential to reveal more true marker-trait associations (Huang et al. 2011).

9.2.6 *Community Resources in Wheat, Barley, Soybean, and Sorghum*

Wheat is a challenging crop in terms of conducting association mapping and genome analysis owing to its large genome size (17 Gb), and difficulties in sequencing and allocating sequences to the A, B, or D genome. Earlier research in wheat has contributed significantly to our understanding of the potential and strategy of association mapping in crops (Brescghello and Sorrells 2006a, b; Sorrells and Yu 2009). Scaling up association mapping in wheat has been hampered mainly by the lack of a sufficiently large number of SNPs (Trebbi et al. 2011); but concerted efforts have been made to sequence the wheat genome by International Wheat Genome Sequencing Consortium (IWGSC) (<http://wheat.pw.usda.gov/PhysicalMapping>) established by plant scientists, plant breeders, and producers to understand the structure and function of the wheat genome (www.wheatgenome.org). This project resulted in the sequencing of wheat genome using 454pyrosequencing, and genome analysis has identified 94,000 to 96,000 genes. This is a resource for accelerating discovery

of genes involved in energy harvesting, metabolism, and phenology of the crop. Wheat genome has many small disruptions to the conserved gene order and is highly dynamic with the loss of several gene family members due to polyploidization and domestication (Brenchley et al. 2012). Beside the efforts towards sequencing the genome, three custom high-throughput SNP genotyping assays (1,536-, 9,000- and 50,000-SNP) based on Illumina BeadArray and Infinium platforms have been developed (Chao et al. 2010) (E Akhunov, personal communication). A set of 2,994 wheat lines were genotyped with the 9,000-SNP iSelect assay (Cavanagh et al. 2013).

To understand the genetic diversity, population structure, and linkage disequilibrium in U.S. elite winter wheat, 205 elite breeding lines were analyzed from U.S. winter wheat breeding programs (Zhang et al. 2010a). The accessions were from the Southern and Northern Regional Performance Nurseries, the Regional Germplasm Observation Nursery, the elite hard winter wheat nursery at Oklahoma State University, the Uniform Eastern Soft Red Winter Wheat Nursery, and the Uniform Southern Soft Red Winter Wheat Nursery, plus 22 major cultivars recently released in the hard winter wheat region. Researchers genotyped the 205 elite breeding lines using 245 SSR markers. Population structure, LD, cluster analysis, and PCA revealed that these collections were highly structured based on their geographical location. The soft and hard winter wheat was separated in the study. The hard winter wheat had more genetic diversity than the soft winter wheat. The LD decay was about 10 cM across the genome (Zhang et al. 2010a).

One of the conclusions from the study was that modern breeding practices maintain reasonable genetic diversity in major U.S. winter wheat breeding gene pools. Also, the presence of higher genetic diversity in hard winter wheat could be used to broaden genetic diversity in soft winter wheat. LD blocks in the genome were identified, but the majority of the genome has lower LD decay, which indicates these could be used for association mapping studies. This study focused on evaluating the germplasm for genetic diversity in the current breeding programs, which will facilitate the use of this information to future cultivar release programs (Zhang et al. 2010a).

Efforts to analyze wheat genetic resources for conducting wheat association studies are progressing. Akhunov et al. performed analysis of nucleotide diversity to construct an SNP database in wheat. They studied about 2,114 genes for nucleotide diversity in *T. aestivum*, *T. dicoccoides* and synthetic 6x wheat. Genetic diversity was similar between A and B genome but was reduced in the D genome. Interestingly, this low variance of the D genome was accompanied by an excess of rare alleles in some genes. The researchers discovered a total of 5,471 SNPs in 1,791 genes. Studying *T. aestivum* and *T. dicoccoides* is a good strategy to develop SNP markers in wheat, where ancestral species are the source of genetic variability. Self pollination and homeologous chromosome pairing could lead to loss of variability in wheat (Akhunov et al. 2010).

Another study on the population structure and genome-wide linkage disequilibrium in wheat used 1536 SNPs (Chao et al. 2010). This study used a panel of 478 spring wheat cultivars from 17 populations across the United States and Mexico, and the population structure analysis identified 9 clusters, indicating that previously

inferred populations share a common genetic identity. The assessment of LD and population structure in this assembled panel of diverse lines provides critical information for the development of genetic resources for genome-wide association mapping of agronomically important traits in wheat (Chao et al. 2010).

In durum wheat (A and B genomes), a sequence-based genotyping approach which couples AFLP-based complexity reduction to next-generation sequencing identified 9,983 putative SNPs between the two parents of a mapping population and used these SNPs to increase map resolution (van Poecke et al. 2013). The same SNPs are being used to profile panels of genotypes suitable for association mapping and characterized by different levels of linkage disequilibrium (M. Maccaferri, personal communication). A panel of elite genotypes has been used to identify QTLs for rust resistance and drought-adaptive traits (Maccaferri et al. 2010, 2011; Letta et al. 2013).

Barley has a large genome (5.1 Gb) and it is an early domesticated crop with high degree of population sub-structure due to breeding activities. More coordinated community-based approaches are followed in crop plants, and barley is no exception. Eight founding institutions from six countries initiated a sequencing project, the International Barley Sequencing Consortium (IBSC) (Schulte et al. 2009), and sequenced the large haploid genome based on a high-resolution genetic map that was anchored using 3.9 Gb of the genome. This consortium identified 79,379 transcript clusters including 26,159 genes (homology support from other plant genome) using whole-genome shotgun assembly, complementary DNA, and deep RNA sequence data. Its genome harbors abundant alternative splicing, premature stop codons, and novel transcriptionally active regions.

Even if the population structure were correlated with the phenotype, by effectively using the statistical methods developed, successful GWAS could be conducted with low marker density. This assumption was validated by successfully mapping 15 morphological traits in barley. Five hundred barley cultivars were genotyped with 1,536 SNPs. The traits studied were seasonal growth (1H), grain lateral nerve speculation (2H), awn anthocyanin coloration, awn anthocyanin intensity, auricle anthocyanin coloration, auricle anthocyanin intensity, lemma nerve anthocyanin intensity (2H), grain aleurone color (4H), hairiness of leaf sheath (4H), rachille hair type (5H), ear attitude (5H), and grain ventral furrow hair (6H) were significantly associated. By developing a double haploid (DH) population from two of the inbreds in the GWAS panel differing in the anthocyanin pigmentation, the *ANT2* gene on chromosome 2H was fine-mapped and validated by genotyping the population using the 1,536 SNP array (Cockram et al. 2010).

GWAS have been conducted to understand the genetic basis of domestication in barley, in which domestication has changed the morphological feature of the inflorescence architecture and resulted in two-rowed and six-rowed forms derived from the ancestor two-row wild types. The development of the six-row barley is controlled by a gene, *VRS1*, on chromosome 2H. But the genome-wide scans show that *INTERMEDIUM-C* located on chromosome 4H is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1 (Tb1)*, which acts as the modifier of the of the *VRS1* gene. Ramsay et al. conducted genome-wide association scans of 190 barley

cultivars by genotyping them with 2,463 biallelic SNPs. This experiment identified three genomic regions on chromosomes 1HL, 2HL, and 4HS associated with the row type. The association on 2H was the *VRS1* gene. The researchers also identified 17 coding mutations in *TBI* correlated with lateral spikelet fertility. The *INT-C* as an ortholog of *ZmTBI* and the confirmation of its involvement in determining both the fertility of the lateral spikelets and of tillering was carried out using a combination of genome-wide association mapping, and studying conservation of synteny and a collection of well-characterized mutant stocks (Ramsay et al. 2011). In barley, a major subdivision is vernalization requirements for flowering in winter and spring. Major flowering time loci (*VRN-H1* and *VRN-H2*) in barley for vernalization requirements were identified through association mapping by studying 429 spring and winter barley accession from Europe (Cockram et al. 2008).

Soybean is an autogamous plant species that exhibits high variation in LD across its genome, indicating that a large number of markers is needed to perform GWAS. The LD pattern in soybean was identified by a study that focused on three genomic regions varying from 336 to 574 kb. The populations used were 26 accessions of the wild ancestor of soybean (*Glycine soja* Seib. et Zucc.), 52 Asian *G. max* landraces, 17 Asian landrace introductions that became the ancestors of North American (N. Am.) cultivars, and 25 elite cultivars from N. Am. In the three cultivated *G. max* groups, LD extended from 90 to 574 kb (Hyten et al. 2007).

Despite high LD in soybean, efforts are ongoing to sequence a number of landraces and cultivars to further understand the genetic structure and to perform association mapping. Seventeen wild and 14 cultivated soybean genomes were sequenced to 5x depth and > 90 % coverage using the Illumina Genome Analyzer II platform. A comparison of the patterns of genetic variation between wild and cultivated soybean identified high allele diversity among the wild soybean. Researchers also identified a set of 205,614 tag SNPs useful for LD mapping and linkage analysis. This is a valuable resource for the analysis of wild soybeans and to facilitate future breeding and quantitative trait analysis.

Large-scale SNP discovery has been conducted by deep resequencing of a reduced representation library (Hyten et al. 2010). Researchers then used the generated SNPs to create a high-resolution map that assisted in the assembly of scaffolds from the 8x whole genome shotgun sequences into pseudomolecules corresponding to soybean chromosomes. As in other crops, the release of the soybean genome sequence (Schmutz et al. 2010) would speed up association mapping related research. Concerted efforts are ongoing to develop a large soybean NAM population with 5,600 lines (B. Diers, personal communication). A set of 40 soybean lines was selected from lines nominated by the soybean community to maximize the genetic diversity based on clustering analysis with 1536 SNPs.

Sorghum is a staple food for the people in sub-Saharan Africa. Its C4 photosynthesis, drought resistance, wide adaptation, and high nutritional value hold the promise to alleviate hunger in Africa. Release of the sorghum genome sequence (Paterson et al. 2009) greatly facilitated research in association mapping. Efforts have been made to assemble information and community resources to conduct association mapping in sorghum and to investigate the LD in sorghum (Casa et al.

2008; Hamblin et al. 2004). Sorghum is an excellent species for association study owing to selfing, low sequence diversity, high LD compared with maize, and availability of a sequenced genome.

A recent SNP discovery project through resequencing 8 diverse sorghum accessions successfully identified 283,000 SNPs (Nelson et al. 2011). This study used the restriction site associated DNA (RAD) approach to construct the sequencing library from only genomic DNA fragments whose 5' ends abut the recognition site of the selected restriction enzymes, *Pst* or *Bsr*FI. SNP discovery rate of the RAD approach was 10-fold higher than that of a semi-random library (digestion by *Hpa*II and fragment size selection of 200–2000 bp).

A panel of 377 lines of sorghum representing major cultivated lines and lines from the sorghum conversion program (SCP) were assembled and characterized for eight traits. Population structure and linkage disequilibrium were estimated (Casa et al. 2008). A 300-line set (a subset of the 377-line set) has also been characterized for grain quality and several candidate genes were identified to harbor SNPs with significant association signals (Sukumaran et al. 2012). A 2000-line sorghum NAM population also has been developed by crossing 10 diverse sorghum lines selected from the sorghum diversity panel with a common parent, Tx430 (Yu et al. 2013).

Candidate gene association mapping has been conducted in sorghum to map the plant height gene. In this study, the sorghum diversity panel was used to effectively characterize the phenotypic effects of the *dw3* mutation and to fine-map a second, epistatic dwarfing QTL on sorghum chromosome 9 (Brown et al. 2008).

In addition, sweet sorghum has the potential to become the crop for bio-energy production. Association mapping has been conducted on sweet sorghum for brix and plant height (Murray et al. 2009). Different sweet sorghum collections were also analyzed for population structure and genetic diversity (Ali et al. 2008; Wang et al. 2009a).

Mostly recently, genome-wide SNP variation has been examined across 971 sorghum accessions with 265,000 SNPs obtained through GBS (Morris et al. 2013). Evidence of selective sweep was found around starch metabolism genes. In addition to mapping several classic plant height loci, GWAS identified known as well as new genes for inflorescence architecture.

9.3 Challenges and Opportunities

Next-generation sequencing technologies provide new opportunities and challenges to the plant genetics communities (Fig. 9.3). New strategies for high-throughput large-scale phenotyping need to be developed to match the level of the genotyping/resequencing capacity. Improved bioinformatics, database, statistical methods, and genetic designs are needed for the large-scale data generated from sequencers, fields, growth chambers, greenhouses, and analytical equipment and scanners (Fig. 9.1). Many new areas will be incorporated into association mapping; for example, genotyping strategies will be tested for detecting copy number variation (CNV)

- | | |
|---|--|
| <input type="checkbox"/> Large scale experiments | <input type="checkbox"/> Genetic Design |
| <input type="checkbox"/> Genome-wide genotyping
CNV, PAV, GBS, RNA-seq | NAM, testcross |
| <input type="checkbox"/> High-throughput phenotyping
Traditional, RNA
Protein/Metabolite
CT-scan, NIR, GPS
Image analysis | <input type="checkbox"/> Validation
RNA-seq
MutMap |
| <input type="checkbox"/> Genetic/statistical Methods
Rare alleles
CR-GWAS
Computational speed
Missing heritability | <input type="checkbox"/> Methylation
Trait, Marker |
| | <input type="checkbox"/> Genomic selection |
| | <input type="checkbox"/> Mapping
Comparative mapping
ShoreMap
Next generation mapping (NGM) |
| | <input type="checkbox"/> Data storage, Analysis, Power |

Fig. 9.3 Future opportunities and challenges that are related to association mapping and the general complex trait dissection and selection

(Rogers and Bendich 1987; Springer et al. 2009) and presence-absence variation (PAV) (Springer et al. 2009). Resequencing strategies using RNA-seq (Wang et al. 2009b), exome sequencing (Ng et al. 2009), and genotyping-by-sequencing (GBS) (Huang et al. 2009; Elshire et al. 2011) are being optimized. More importantly, the recent sequencing effort in wheat (Brenchley et al. 2012) and barley (Mayer et al. 2012) genomes provides an excellent resource and opportunity to speed up gene discovery for the improvement of crop plants along with the high-throughput phenotyping.

High-throughput phenotyping is likely to be the most expensive part of plant genetic studies. Phenotypes used in association mapping are being expanded from the traditional, labor-intensive phenotyping to gene expression, and protein/metabolite level. New techniques such as CT-SCAN, near infrared (NIR) spectroscopy, single kernel characterization system for grain quality, global positioning system, and image analysis are all being exploited for faster and more accurate phenotyping. More studies focusing on rare alleles, multiple alleles, computational speed, data storage, and computational power are also needed. Validation of the results using RNA-seq and other approaches will be important. Methylation in the genome could be used as a marker (Laird 2003) and or as a trait (Lukens and Zhan 2007). Research in association mapping will certainly be integrated with other areas such as MutMap (Abe et al. 2012), next-generation mapping (NGM) (Schneeberger et al. 2009), genomic selection (Bernardo and Yu 2007; Meuwissen et al. 2001), and comparative genomics.

9.3.1 *Missing Heritability*

Association mapping strategy is based on the assumption that the common phenotypic variation are caused by common genetic variants. As a result, array-based genotyping has been used extensively in human GWAS because these SNPs were selected to represent the major genetic polymorphisms and were expected to explain a decent proportion of phenotypic variation of the trait. However, if we summarize the effect of genes identified through GWAS and they explain only a small proportion of the phenotypic variation. This turns out to be the case for a series of human complex traits. At the same time, heritability estimates at the population level are adequately high. Consequently, this discrepancy between what we expected to find and what we found was termed as the “missing heritability” issue in human GWAS (Manolio et al. 2009).

The causes of missing heritability can be attributed to a number of factors: genotype by environment interaction, a larger number of variants of smaller effects yet to be found; rare variants (possibly with larger effects) that are poorly detected by available genotyping arrays that focus on variants present in 5 % or more of the population; structural variants poorly captured by existing arrays; low power to detect gene—gene interactions, inadequate accounting for shared environment among relatives, statistical issues, copy number variation (CNV), and multiple testing issues. The power of GWAS to detect variants of modest effect and low frequency remains lacking due to the low frequency of functional alleles in the mapping population, the low influence of low-frequency alleles on the population, and/or lower detection power of the association mapping strategy. The phenotypic variation caused by numerous small-effect alleles will be difficult to detect compared with a small number of large-effect alleles; this is a challenge for any complex trait dissection studies, including association mapping.

9.3.2 *New Gene Identification*

Identifying previously unknown genes underlying a complex trait remains to be challenging. Once association signals are detected, concerted efforts are required to identify the causal genes or polymorphisms. These follow-up studies need to be well designed given the complex genome of major crops and difficulties in choosing appropriate genetic backgrounds for genetic and transformation validation. If the complete genome sequence is unavailable while conducting association mapping, new genes with small to modest effects are difficult to be recognized and followed up in further studies. On the other hand, the candidate gene approach by definition is not designed to identify novel genes underlying a complex trait. In addition, allele frequency, multiple testing, and epistasis add to the problem of low detection rate. Compared with human genetics, the number of GWAS in plants with adequate sample size and marker density is very limited. However, advances in sequencing and GWAS methodology, completion of draft genome sequence or even multiple

reference genome sequences, and continued improvements in different genetic and genomic techniques would eventually make it possible to realize the potential offered by association mapping in identifying new genes underlying complex traits.

9.3.3 Genotyping-by-Sequencing (GBS)

Rapidly evolving sequencing and genotyping technologies have fundamentally changed not only the design of specific breeding and selection strategies in crops, but also how the vast amount of available germplasm diversity can be utilized efficiently (Bernardo and Yu 2007; Heffner et al. 2009; Tester and Langridge 2010). Routine use of GS in plant breeding is becoming possible because of the significantly reduced cost of obtaining molecular marker information, particularly SNPs, thanks to the development of high-throughput technology from DNA extraction, sample preparation, and array-based genotyping technologies as well as cutting-edge GBS technology (Metzker 2010). Current GBS research includes species with a sequenced genome, such as rice (Huang et al. 2009), maize (Elshire et al. 2011), and sorghum (Morris et al. 2013), and those without, such as wheat and barley (Chutimanitsakun et al. 2011). Next-generation sequencing technologies have improved output and made possible sequencing of multiple samples at the same time. Sequencing-based high-throughput genotyping combines the advantages of cost-effectiveness, less time, and dense marker data. In the first GBS paper (Huang et al. 2009), a sliding window approach for analyzing the SNPs collectively rather than individually was used on 150 RILs derived from the cross between *indica* and *japonica* rice cultivars. The SNP calling in this method is based on a recombination break point and sliding window. With the sequence-based genetic map, a 100-kb region was identified for plant height that is related to a green revolution gene (Huang et al. 2009).

The key step in conducting GBS is to reduce the genome complexity through restriction enzymes (REs) digestion, and this is particularly important given the high proportion of repeats in many crop genomes. Methylation-sensitive REs are used to reduce the genome complexity so that lower copy regions are targeted with higher efficiency. This method also simplifies the challenges of sequence alignment problems (Elshire et al. 2011). A two-enzyme digestion protocol was also developed for barley and wheat (Poland et al. 2012) and a detailed review of progress in GBS and its application were published (Poland and Rife 2012).

9.3.4 Rare Alleles

At present, GWAS is unable to detect rare variants through common SNP markers (Ott et al. 2011). The power to detect an association is a function of allele frequency, and individually, rare alleles have little influence on the population, which renders their detection difficult. The difficulty in detecting the rare alleles is more

of a biological problem than a statistical issue. A Composite Resequencing-Based Genome-Wide Association Studies (CR-GWAS) approach was recently proposed to address this issue (Zhu et al. 2011). This approach integrates next-generation sequencing, prediction of biological function of SNPs, statistical test for rare allele variants, and genome databases and gene networks.

9.3.5 *Genic and Nongenic Contribution*

Notably, signals generated from GWAS can be tabulated to characterize the genome-wide landscape of genetic polymorphisms underlying quantitative trait variation (Li et al. 2012). Unlike previous analyses, a set of quantitative traits were systematically analyzed with a two-stage genome scan method to quantify the contributions from genic and nongenic SNPs. Trait-associated SNPs (TASs) for these quantitative traits were found to be enriched in non-genic regions, particularly within a 5-kb window upstream of genes. Besides highlighting the importance of polymorphisms regulating gene expression in shaping the natural variation, these findings also suggested that efficient, cost-effective routine GWAS in species with complex genomes can focus on genic and promoter regions.

In summary, association mapping has become one of the major approaches in gene discovery and complex trait dissection in plants. Combined with advanced genetic designs in plant genetics, newer algorithms for speedy analysis of the data and decision making, and the development of immortal populations such as NAM showcased the potential of what can be achieved by assimilating knowledge and discovery in other research areas. Utilization of the rapid advancement in next generation sequencing technologies is expected to foster yield advantages in all major crops. Concerted efforts are ongoing in almost all major plant species, and we expect findings from these studies would ultimately enrich our understanding and foster the development of more efficient approaches.

Acknowledgments This work was supported by the Agriculture and Food Research Initiative Competitive Grant (2011-03587) from the USDA National Institute of Food and Agriculture, the Plant Feedstock Genomics Program (DE-SC0002259) of the U.S. Department of Energy, the Plant Genome Program (DBI-0820610) of the National Science Foundation, the Targeted Excellence Program of Kansas State University, and the Kansas State University Center for Sorghum Improvement.

References

- Abe A, Kosugi S, Yoshida K et al (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotechnol* 30:174–178
- Akhunov E, Akhunova A, Anderson O et al (2010) Nucleotide diversity maps reveal variation in diversity among wheat genomes and chromosomes. *BMC Genomics* 11:702

- Ali M, Rajewski J, Baenziger P et al (2008) Assessment of genetic diversity and relationship among a collection of US sweet sorghum germplasm by SSR markers. *Mol Breed* 21:497–509
- Atwell S, Huang YS, Vilhjalmsón BJ et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627–631
- Aulchenko YS, de Koning DJ, Haley C (2007) Genome wide rapid association using mixed model and regression: a fast and simple method for genome wide pedigree-based quantitative trait loci association analysis. *Genetics* 177:577–585
- Bernardo R, Yu J (2007) Prospects for genome-wide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090
- Brachi B, Faure N, Horton M et al (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet* 6:e1000940
- Bradbury PJ, Zhang Z, Kroon DE et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 23:2633–2635
- Brenchley RC, Spannagl M, Pfeifer M et al (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705–710
- Breseghello F, Sorrells ME (2006a) Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci* 46:1323–1330
- Breseghello F, Sorrells ME (2006b) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172:1165–1177
- Brown PJ, Rooney WL, Franks C, Kresovich S (2008) Efficient mapping of plant height quantitative trait loci in a sorghum association population with introgressed dwarfing genes. *Genetics* 180:629–637
- Browning B, Browning S (2008) Haplotypic analysis of wellcome trust case control consortium data. *Hum Genet* 123:273–280
- Buckler ES, Gaut BS, McMullen MD (2006) Molecular and functional diversity of maize. *Curr Opin Plant Biol* 9:172–176
- Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
- Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172:557–567
- Camus-Kulandaivelu L, Veyrieras JB, Madur D et al (2006) Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* 172:2449–2463
- Cao J, Schneeberger K, Ossowski S et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat Genet* 43:956–963
- Casa AM, Pressoir G, Brown PJ et al (2008) Community resources and strategies for association mapping in sorghum. *Crop Sci* 48:30–40
- Cavanagh CR, Chao S, Wang S, Huang BE, Stephen S, Kiani S et al (2013) Genome-wide comparative diversity uncovers multiple targets of selection for improvement in hexaploid wheat landraces and cultivars. *Proc Natl Acad Sci USA* 110:8057–8062
- Chao S, Dubcovsky J, Dvorak J et al (2010) Population- and genome-specific patterns of linkage disequilibrium and SNP variation in spring and winter wheat (*Triticum aestivum* L.). *BMC Genomics* 11:727
- Chia JM, Song C, Bradbury PJ et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat Genet* 44:803–807
- Chutimanitsakun Y, Nipper RW, Cuesta-Marcos A et al (2011) Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. *BMC Genomics* 12:4
- Clark RM, Schweikert G, Toomajian C et al (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–342
- Cockram J, White J, Leigh F et al (2008) Association mapping of partitioning loci in barley. *BMC Genet* 9:16
- Cockram J, White J, Zuluaga DL et al (2010) Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proc Natl Acad Sci USA* 107:21611–21616

- Crainiceanu CM, Ruppert D (2004) Likelihood ratio tests in linear mixed models with one variance component. *J R Stat Soc B* 66:165–185
- Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997–1004
- Devlin B, Roeder K, Wasserman L (2001) Genomic Control, a New Approach to Genetic-Based Association Studies. *Theor Popul Biol* 60:155–166
- Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379
- Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. *Ann Rev Plant Biol* 54:357–374
- Flint-Garcia SA, Thuillet AC, Yu J et al (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* 44:1054–1064
- Gilmour AR, Gogel BJ, Cullis BR et al (2002) ASReml user guide release 1.0. VSN International Ltd, Hemel Hempstead
- Gore MA, Chia J-M, Elshire RJ et al (2009) A first-generation haplotype map of maize. *Science* 326:1115–1117
- Hamblin MT, Mitchell SE, White GM et al (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. *Genetics* 167:471–483
- Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Mol Ecol Notes* 2:618–620
- Harjes CE, Rocheford TR, Bai L et al (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* 319:330–333
- Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
- Henderson CR (1975) Comparison of alternative sire evaluation methods. *J Anim Sci* 41:760–770
- Huang X, Feng Q, Qian Q et al (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res* 19:1068–1076
- Huang X, Wei X, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–967
- Huang X, Zhao Y, Wei X et al (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* 44:32–39
- Huang X, Kurata N, Wei X et al (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497–501
- Hyten DL, Cannon SB, Song Q et al (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. *BMC Genomics* 11:38
- Hyten DL, Choi I-Y, Song Q, Shoemaker RC (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175:1937–1944
- Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–1723
- Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–354
- Kariya T, Kurata H (2004) Generalized least squares estimators. *Generalized least squares*. John Wiley & Sons Ltd, pp 25–66
- Konishi S, Izawa T, Lin SY et al (2006) An SNP caused loss of seed shattering during rice domestication. *science* 312:1392–1396
- Kump KL, Bradbury PJ, Wissler RJ et al (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat Genet* 43:163–168
- Laird PW (2003) The power and the promise of DNA methylation markers. *Nat Rev Cancer* 3:253–266
- Letta T, Maccaferri M, Badebo A et al (2013) Searching for novel sources of field resistance to Ug99 and Ethiopian stem rust races in durum wheat via association mapping. *Theor Appl Genet* 126:1237–1256

- Lipka AE, Tian F, Wang Q et al (2012) GAPIT: genome association and prediction integrated tool. *Bioinformatics* 28:2397–2399
- Li X, Zhu C, Yeh C-T et al (2012) Genic and non-genic contributions to natural variation of quantitative traits in maize. *Genome Res* 22:2436–2444
- Lippert C, Listgarten J, Liu Y et al (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* 8:833–835
- Liu KJ, Goodman M, Muse S et al (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165:2117–2128
- Lu H-Y, Liu X-F, Wei S-P, Zhang Y-M (2011) Epistatic association mapping in homozygous crop cultivars. *PLoS ONE* 6:e17773
- Lukens LN, Zhan S (2007) The plant genome's methylation status and response to stress: implications for plant improvement. *Curr Opin Plant Biol* 10:317–322
- Maccaferri M, Sanguineti MC, Tuberosa R (2005) Analysis of linkage disequilibrium in a collection of elite durum wheat genotypes. *Mol Breed* 15:271–289
- Maccaferri M, Sanguineti MC, Natoli E et al (2006) A panel of elite accessions of durum wheat (*Triticum durum* Desf.) suitable for association mapping studies. *Plant Genet Res* 4:79–85
- Maccaferri M, Sanguineti MC, Mantovani P et al (2010) Association mapping of leaf rust response in durum wheat. *Mol Breed* 26:189–228
- Maccaferri M, Sanguineti MC, Demontis A et al (2011) Association mapping in durum wheat grown across a broad range of water regimes. *J Exp Botany* 62:409–438
- Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
- Mayer KF, Waugh R, Brown JW et al (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491:711–716
- McMullen MD, Kresovich S, Villeda HS et al (2009) Genetic properties of the maize nested association mapping population. *Science* 325:737–740
- Morris GP, Ramu P, Deshpande SP et al (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc Natl Acad Sci USA* 110:453–458
- McNally KL, Bruskiewich R, Mackill D et al (2006) Sequencing multiple and diverse rice varieties. Connecting whole-genome variation with phenotypes. *Plant Physiol* 141:26–31
- McNally KL, Childs KL, Bohnert R et al (2009) Genome-wide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci USA* 106:12273–12278
- Metzker ML (2010) Applications of next-generation sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Murray SC, Rooney WL, Hamblin MT et al (2009) Sweet sorghum genetic diversity and association mapping for brix and height. *Plant Genome* 2:48–62
- Myles S, Peiffer J, Brown PJ et al (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202
- Nelson JC, Wang S, Wu Y et al (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. *BMC Genomics* 12:352
- Ng SB, Turner EH, Robertson PD et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276
- Nordborg M, Hu TT, Ishino Y et al (2005) The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biol* 3:e196
- Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. *Nat Rev Genet* 12:465–474
- Palaisa KA, Morgante M, Williams M, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two *phytoene* synthase loci. *Plant Cell* 15:1795–1806
- Paterson AH, Bowers JE, Bruggmann et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–556

- Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2:e190
- Platt A, Horton M, Huang YS et al (2010) The scale of population structure in *Arabidopsis thaliana*. *PLoS Genet* 6:e1000843
- Poland JA, Bradbury PJ, Buckler ES (2010) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc Natl Acad Sci USA* 108:6893–6898
- Poland JA, Brown PJ, Sorrells ME et al (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* 7:e32253
- Poland JA, Rife TW (2012) Genotyping-by-Sequencing for plant breeding and genetics. *Plant Genome* (in press)
- Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
- Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) association mapping in structured populations. *Am J Hum Genet* 67:170–181
- Quaas RL, Pollak EJ (1981) Modified equations for sire models with groups. *J Dairy Sci* 64:1868–1872
- Ramsay L, Comadran J, Druka A et al (2011) *INTERMEDIUM-C*, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1*. *Nat Genet* 43:169–172
- Remington DL, Thornsberry JM, Matsuoka Y et al (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 98:11479–11484
- Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
- Rogers SO, Bendich AJ (1987) Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. *Plant Mol Biol* 9:509–520
- Salvi S, Sponza G, Morgante M et al (2007) Conserved non-coding genomic sequences controlling flowering time differences in maize. *Proc Natl Acad Sci USA* 104:11376–11381
- Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183
- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Schneeberger K, Ossowski S, Lanz C et al (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Meth* 6:550–551
- Schulte D, Close TJ, Graner A et al (2009) The international barley sequencing consortium-at the threshold of efficient access to the barley genome. *Plant Physiol* 149:142–147
- Segura V, Vilhjálmsson BJ, Platt A et al (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* 44:825–830
- Sequencing Project International Rice G (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Shomura A, Izawa T, Ebana K et al (2008) Deletion in a gene associated with grain size increased yields during rice domestication. *Nat Genet* 40:1023–1028
- Sorrells ME, Yu J (2009) Linkage disequilibrium and association mapping in the *Triticeae*. In: Feuillet C, J. MG (eds) *Genetics and genomics of the triticeae*, plant Genetics/Genomics. Springer Verlag, pp 655–684
- Springer NM, Ying K, Fu Y et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLoS Genet* 5:e1000734
- Sukumaran S, Xiang W, Bean SR et al (2012) Association mapping for grain quality in a diverse sorghum collection. *Plant Genome* 5:126–135
- Tester M, Langridge P (2010) Breeding technologies to increase crop production in a changing world. *Science* 327:818–822
- Thornsberry JM, Goodman MM, Doebley J et al (2001) *Dwarf8* polymorphisms associate with variation in flowering time. *Nat Genet* 28:286–289

- Tian F, Bradbury PJ, Brown PJ et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
- Trebbi D, Maccaferri M, de Heer P et al (2011) High-throughput SNP discovery and genotyping in durum wheat (*Triticum durum* Desf.). *Theor Appl Genet* 123:555–569
- van Poecke RMP, Maccaferri M, Tang J et al (2013). Sequence-based SNP genotyping in durum wheat (submitted).
- Wang M, Zhu C, Barkley N et al (2009a) Genetic diversity and population structure analysis of accessions in the US historic sweet sorghum collection. *Theor Appl Genet* 120:13–23
- Wang Z, Gerstein M, Snyder M (2009b) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63
- Weber A, Clark RM, Vaughn L et al (2007) Major regulatory genes in maize contribute to standing variation in *Teosinte* (*Zea mays* ssp. *parviglumis*). *Genetics* 177:2349–2359
- Weigel D, Mott R (2009) The 1001 Genomes project for *Arabidopsis thaliana*. *Genome Biol* 10:107
- Wisser RJ, Kolkman JM, Patzoldt ME et al (2011) Multivariate analysis of maize disease resistances suggests a pleiotropic genetic basis and implicates a *GST* gene. *Proc Natl Acad Sci USA* 108:7339–7344
- Yan J, Kandianis CB, Harjes CE et al (2010) Rare genetic variation at *Zea mays crtR1* increases β -carotene in maize grain. *Nat Genet* 42:322–327
- Yu J, Arbelvide M, Bernardo R (2005) Power of in silico QTL mapping from phenotypic, pedigree, and marker data in a hybrid breeding program. *Theor Appl Genet* 110:1061–1067
- Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* 17:155–160
- Yu J, Hamblin MT, Tuinstra MR (2013) Association Genetics Strategies and Resources. In: Paterson A (ed) genetics and genomics of the *Saccharinae*. *Plant Genet Genom: crops and models* 11:187–203
- Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539–551
- Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhang D, Bai G, Zhu C et al (2010a) Genetic diversity, population structure, and linkage disequilibrium in U.S. elite winter wheat. *Plant Genome* 3:117–127
- Zhang Z, Ersoz E, Lai C-Q et al (2010b) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42:355–360
- Zhao K, Aranzana MJ, Kim S et al (2007) An *Arabidopsis* example of association mapping in structured samples. *PLoS Genet* 3:e4
- Zhao K, Tung C-W, Eizenga GC et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* 2:467
- Zhao K, Wright M, Kimball J et al (2010) Genomic diversity and introgression in *O. sativa* reveal the impact of domestication and breeding on the rice genome. *PLoS ONE* 5:e10780
- Zhou X, Stephens M (2012) Genome-wide efficient mixed model analysis for association studies. *Nat Genet* 44:821–824
- Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. *Plant Genome* 1:5–20
- Zhu C, Li X, Yu J (2011) Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CR-GWAS). *G3: Genes, Genomes*. *Genetics* 1:233–243
- Zhu C, Yu J (2009) Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* 182:875–888

Chapter 10

Exploiting Barley Genetic Resources for Genome Wide Association Scans (GWAS)

Robbie Waugh, Andrew J. Flavell, Joanne Russell, William (Bill) Thomas, Luke Ramsay and Jordi Comadran

Contents

10.1	Introduction	238
10.2	Multi Parent Populations	239
10.3	Linkage Disequilibrium	239
10.4	Population Structure	241
10.5	Genetic Markers	244
10.6	Ascertainment Bias	245
10.7	GWAS	247
10.8	Future Prospects	250
	References	251

Abstract We have been exploring the use of GWAS for trait analysis and gene isolation in cultivated barley. In this chapter we describe the approach we have taken and some of the hurdles that we have faced when attempting to establish the whole system. We discuss the way that we, but also others, have addressed the various issues that have arisen and provide guidance on how they can be avoided. These range from choosing the appropriate population for analysis, how to deal with inherent population structure, genetic marker discovery, application and the effect of ascertainment bias to the range of software currently available for conducting association analyses. We conclude by providing a series of successful examples from our laboratory that range from analysis of simple single gene traits through oligogenic to quantitative traits, and the detection of epistatic interactions. We conclude that appropriately designed and executed GWAS in barley is a powerful tool in our quest to identify the genes and alleles underlying key genetic traits.

R. Waugh (✉) · J. Russell · W. (Bill) Thomas · L. Ramsay · J. Comadran
The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, Scotland
e-mail: Robbie.Waugh@hutton.ac.uk

A. J. Flavell
Division of Plant Sciences, The University of Dundee at JHI,
Invergowrie, Dundee, DD2 5DA, Scotland

Keywords Barley · Germplasm · Linkage disequilibrium · Association mapping

10.1 Introduction

Crop plants evolved from their wild ancestors by the processes of domestication and selective breeding over the last *ca.* 10,000 years. Initially, wild plants carrying promising traits were cultivated, leading eventually to locally adapted landraces. These lost many undesirable alleles as useful alleles became enriched (Feuillet et al. 2008). Modern breeding has largely extended this by a process of crossing the ‘best with the best’ and the successes have been impressive. Unfortunately, there are indications that we are approaching a performance ceiling for at least some crops, as the best alleles become assembled in elite genetic materials (Tanksley and McCouch 1997, <http://www.fao.org/ag/agp/agpc/doc/riceinfo/Asia/ASIABODY.HTM>). The potential to re-invigorate these elite materials may be provided by the introduction of new alleles from wild species and old, locally adapted germplasm. Many studies have demonstrated the value of alleles originating from un-adapted and unimproved germplasm showing that centuries of selective breeding have not necessarily resulted in the accumulation of all the optimal alleles. For example, several barley cultivars have been released in Europe that contain fungal resistance genes introgressed recently from *H. spontaneum* (von Korff et al. 2005; Schmalenbach et al. 2009). A major challenge for the future is to streamline this process using high throughput genomics approaches.

The identification and recruitment of useful alleles are two very different tasks and both are difficult. Allele identification requires detailed and careful phenotypic trait analysis, combined with high-resolution genomic characterisation. Comparison between the phenotypic and genotypic data sets, either by linkage mapping in biparental populations or by genome wide association scanning (GWAS) of panels of related genotypes can in principle yield candidate marker alleles linked to the traits investigated. While the former approach has been generally successful in identification, deployment of the results in breeding has not been as widespread for many reasons, including the problems in identifying markers sufficiently closely linked for effective use in selection. The latter approach is therefore becoming more attractive because it is intrinsically higher resolution and, has the potential at least, to be more powerful because it scrutinises the results of many more generations of recombination and selection (Caldwell et al. 2006; Rostoks et al. 2006; Cockram et al. 2010). However there are also issues with GWAS that need to be resolved before it can be most effectively applied. In this chapter we will review some of the challenges that we have encountered and that need to be considered when planning to exploit genetic resources for GWAS. These are largely based on our experiences in establishing a successful GWAS programme in barley.

10.2 Multi Parent Populations

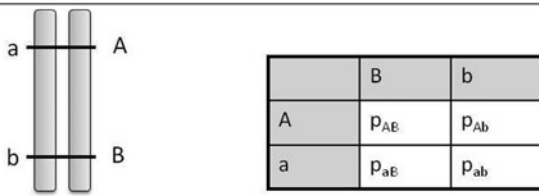
Over the past 25 years the correlation of phenotypic data with genetic markers in the offspring from specific bi-parental crosses using the well-established methods of ‘genetic linkage analysis’ has significantly advanced our understanding of the number, organization, location, and contribution of genetic loci to both simple and complex phenotypes (e.g. Turner et al. 2005; Yan et al. 2006). In a growing number of cases, particularly for Mendelian (i.e. single gene) traits, linkage mapping in very large populations has allowed the responsible genes to be fine mapped and ultimately to be cloned and analysed at the sequence level (Komatsuda et al. 2007). This has been achievable because the large number of recombination events in such populations allows the trait gene to be positioned so accurately that it is often possible to resolve its location to a specific DNA sequence (when available) or a single large-insert DNA clone that contains only one or perhaps a few candidate genes. Successes include major disease resistance and developmental genes such as *Mlo*, *Rpg1*, *Vrn1* and *Ppd1* and more will continue into the future. Bi-parental mapping requires the construction of specific populations that segregate for the trait of interest and because it samples only a small portion of the genetic variation inherent in the genepool under study, different populations are frequently required for each new trait studied.

More recently, geneticists have started to investigate GWAS in an attempt to increase the resolution of primary genetic studies. In contrast to linkage analysis, association approaches evaluate the correlation between loci and/or markers in populations of plants that share a degree of common history. Populations used for GWAS include collections of related individuals within natural or constructed populations from within a species. Association mapping effectively increases the number of recombination events to include all occurrences within the history of the sample. This presents a distinct advantage over bi-parental populations by improving genetic resolution from the megabase to the kilobase scale. The resolution inherent in a population used for GWAS is largely dependent upon the phenomenon of linkage disequilibrium a measure that can itself be complicated by the history of the population and which has the potential to increase the frequency false positive associations.

10.3 Linkage Disequilibrium

Linkage Disequilibrium (LD) is defined as the non-independence of alleles at different loci in a population (Box 1). At its most basic level, LD is maintained as a balance between mutation and recombination. At the moment of spontaneous (or induced) generation all new mutations are in perfect association with their genetic background. However, over time the processes of recombination (during meiosis) and genetic drift gradually lead to decay in the extent of these original associations and as new mutations are generated and selected, and old ones are lost, new associations are established. LD is therefore the product of evolutionary and biological factors that together contribute to the genetic structure and allelic histories of each

Box 1



For 2 loci each with two alleles, A and a at the first locus and B and b at the second, LD between these loci is given by:

$$R^2 = (p_{AB}p_{ab} - p_{Ab}p_{aB})^2 / p_A p_a p_B p_b$$

Where:
 p_A etc is the frequency of allele A in the population
 p_{AB} is the frequency of individuals with A allele at first locus and B allele at the second locus

Note:
 R^2 measures statistical association and there is a simple inverse relationship between this measure and the sample size.
 R^2 takes a value of 1 if only two haplotypes are present.

gene in the population. The extent of LD can be measured effectively by assaying and correlating the allelic state of genetically linked molecular markers at known genetic loci across the genome in what has been termed an association mapping panel of genotypes. When LD is extensive, statistically significant associations (correlations) may be detected between markers that are several to many centi-Morgans (i.e. potentially several megabases) apart. When it is low, associations between genes or markers may rapidly reduce to become non-significant at the sub-centiMorgan scale, or over thousands or even hundreds of bases. Within this generalised assertion, false positive associations can arise from the effects of genetic structure in the population, which may have originated from non-random mating, population bottlenecks or directional selection. As an example, up to 80% of the significant associations detected between polymorphisms in the maize *dwarf8* (*d8*) gene and flowering time were assessed as being due to population substructure (Thornsberry et al. 2001).

Mating system has a similarly profound impact on LD. Simulation studies have demonstrated that in the absence of mitigating factors, high levels of LD persist to a greater extent in highly selfing species (like barley), and that this is predominantly a factor of the *effective* recombination rate. This is simply because inbreeding results in increased homozygosity. Subsequently, as a consequence of this high homozygosity, a significant proportion of all recombination events in an inbreeding species will fail to bring about an exchange of genetic variation. This has been countered to some extent by artificial outcrossing, the basis of plant breeding practiced over the last hundred years. Therefore, in inbreeding crops like barley, while we would naturally

expect LD to be extensive in natural populations, plant breeding has been effective at generating a pseudo-outcrossing population where LD has been reduced to an extent that makes it useful for medium resolution association-based approaches and the identification of correlations between trait genes and alleles at molecular marker loci (Rostoks et al. 2006).

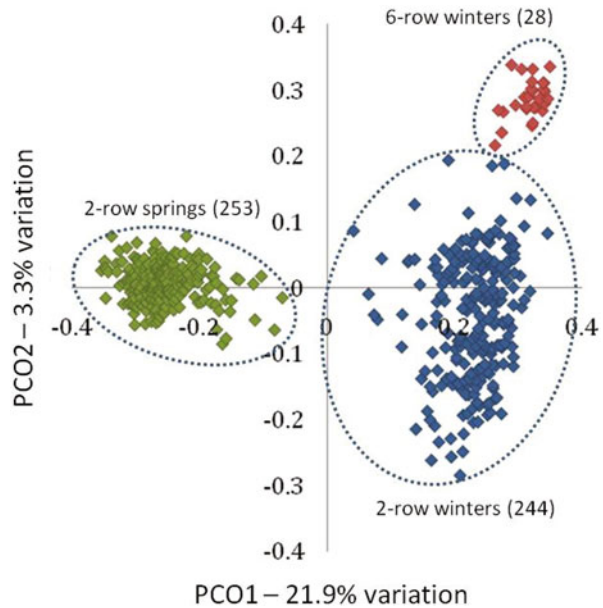
While it is relatively easy to detect marker-trait associations if there is extensive LD this inevitably results in a lower resolution map that requires more work to pin down the allele associated with the trait under study. Natural populations (including both true wild plants and adapted cultivated landraces) contain high levels of genetic diversity and are a great potential reservoir of DNA variation for crop improvement. Because of their history (i.e. number of generations), they also exhibit less extensive LD (Morrell et al. 2005; Kraakman 2005; Caldwell et al. 2006). These are potentially valuable as populations with low LD provide an opportunity to reveal high-resolution associations. Of course, if a genome wide approach is being adopted, the number of markers needed to find any associations would need to be extremely high, which is an associated cost. This has led to the suggestion that, at least in principle, associations could be mapped to an approximate genomic location in germplasm where LD is extensive, then exact genomic regions could be saturated using progressively wider germplasm with correspondingly lower LD but higher marker densities around the established location of the causal gene. In practice this has not yet been achieved.

10.4 Population Structure

For association mapping, the underlying population structure can be a strong confounding factor, especially for traits that have driven the geographical or environmental adaptation of the germplasm set. From a practical point of view, considerable care therefore has to be taken in choosing germplasm, avoiding—if possible—the inclusion of strong population stratification given it is a source of false positive associations. In other words, for a specific trait if there were major loci associated with genetically distinct homogeneous clusters of lines, many background markers carrying alleles exclusive to the specific clusters are also going to be associated with the trait, even though they are not causal. Not surprisingly, a number of approaches have been used to minimise these effects.

Statistical Approaches Our genome-wide association mapping studies in barley (*Hordeum vulgare*) have forced us to confront the problem of population structure as a confounding factor. Barley germplasm is strongly stratified reflecting crop type (in terms of growth habit and spike morphology) and geographical origin, which is heavily linked to local adaptation of the germplasm (Fig. 10.1). For most studies, genotyping and phenotyping are conducted simultaneously. Thus, the exploration and statistical adjustment for stratification is generally conducted within the running time of a project and there is little scope for choosing a different set of lines if structure turns out to be a considerable problem. Moreover, after expensive and time-consuming data collection, a natural tendency is to want include as many data points

Fig. 10.1 Population structure in the cultivated elite barley gene pool (523 lines with 890 non position-redundant SNPs). Three main clusters are evident based on the major biological divisions within the species



as possible in an analysis. Thus statistical approaches that correct and/or account for the effects of population structure within association scans have guided most of the research on GWAS for the last few years. Several different approaches have been proposed in the literature (Mackay and Powell 2007). Issues however arise when the number and identity of markers that remain significant after employing different statistical population structure correction methods are either inconsistent or remove known biological factors correlated at some level with the population stratification. This can result in uncertainty over what QTL to prioritise for further studies or to use as diagnostics in Marker Assisted Selection (MAS).

It is worth mentioning that in an association panel the ancestral marker allele frequencies are not known. Therefore even with saturated genome coverage, is it not possible to build a genetic map *de novo* using LD and then to use this as a framework for visualizing the location of QTL. Thus, a prior genetic map using one or several bi-parental populations needs to be built in parallel to the association mapping panel to estimate the genetic, or better physical, order of the markers in the genome, unless of course the genome sequence of the target species has been assembled. Some of the main approaches for dealing with structure are:

Structured Association Structured association uses multiple polymorphisms assayed throughout the genome to compute statistics that capture the underlying population structure of the germplasm—introducing non-independence between genotypes as a result of common genetic background. Statistics can be then modelled within a Mixed Linear Model (MLM) framework to account for multiple levels of relatedness due to historical population structure and kinship (Yu et al. 2006). Different

software/ statistical packages—for example R v 2.9.0 (<http://www.R-project.org/>), TASSEL v.3.0 (<http://www.maizegenetics.net>) or Genstat 14 (VSN International 2011)—provide different ways of correcting for population structure which can be used to assess which best suits your data. A variance covariance matrix containing coefficients of co-ancestry (kinship matrix) can be included in the mixed model to account for genetic relatedness between genotypes. Eigenanalysis uses the scores of the most significant principal components from the molecular marker matrix as co-variables in the mixed model, which is an approximation to the use of a kinship matrix. In barley, we found a mixed linear regression model (Yu et al. 2006), which accounts for multiple levels of relatedness due to historical population substructure and kinship, to perform best either implemented on its own and in combination with other methodologies. The significance threshold is usually estimated for each analysis using a Bonferroni corrected p -value of 0.05.

With the rapid increase of the amount of SNP marker data there is a need for methods that are able to cope with thousands to millions of computationally intensive analyses. To deal with this, emerging methodologies provide us with a choice of both approximate [e.g. GRAMMAR (Aulchenko et al. 2007), implemented in GenABEL (<http://www.genabel.org/packages/GenABEL>), P3D (Zhang et al. 2010), implemented in TASSEL (<http://www.maizegenetics.net/tassel>), EMMAX (Kang et al. 2010) (<http://genetics.cs.ucla.edu/emmax/>)] and exact methods [e.g. FMM (W. Astle & D. Balding, <http://www.genabel.org/MixABEL/FastMixedModel.html>), FaST-LMM (Lippert et al. 2011) (<http://mscompbio.codeplex.com/>), GEMMA [M. Stephens lab (<http://stephenslab.uchicago.edu/software.html>)] to account for structure effects.

Naive Approach In its simplest form, the *naive* approach—which does not account for any population structure correction—is based on the same principles that work for bi-parental QTL mapping populations and consists of a regression of the phenotype upon the genotype to detect the QTLs. Each marker in a genetic map has a probability to be associated with the QTL of interest. The naive approach is suitable for use in the following two types of population—though some would argue that as all populations have some residual structure, a structure correction should always be applied.

Constructed Populations New population types that capture the advantages of both linkage mapping and GWAS, and that focus on achieving high statistical power, high resolution and low population stratification have been developed in several species and have, or are, being developed in barley. Nested Association Mapping (NAM) (McMullen et al. 2009) and heterogenic stock inbred lines, also known as multi-parent advanced generation intercross or MAGIC populations overcome the handicaps imposed by stratification in natural germplasm collections (Cavanagh et al. 2008). Trait mapping using NAM and MAGIC populations is more complete due to greater genetic diversity and more precise than classical bi-parental populations. The short history of recombination gives high statistical power to QTL detection, while ancestral recombination and diversity accumulated between the parental lines provide the basis for much finer scale mapping. Rounds of inter-crossing and selfing remove long range LD present between the parental lines, and each extra generation

will shuffle the genetic contribution from the founder lines more and more. For NAM in Maize, twenty-five diverse lines were crossed to B73 and the F1 plants self-fertilized for six generations to create a series of twenty-five recombinant inbred line (RIL) families ultimately totalling 5000 individuals. In MAGIC populations a complex and time-consuming crossing scheme has to be implemented to avoid the creation of clusters of highly related progenies that could potentially introduce *de novo* germplasm stratification.

Sub-Populations Artificial out-crossing imposed by breeders coupled with the long recombination history of crop germplasm can create a highly diverse germplasm stock without major population sub-divisions. Assembling a population of this type is the approach we have taken. By exploiting the European elite two-rowed spring barley genepool, our association mapping population effectively behaves like a heterogenic stock inbred line population without strong stratification. It lacks confounding population effects and its assembly avoided complex and time-consuming crossing schemes. Most important from our point of view was that it enabled us to perform QTL analysis and discovery in a germplasm set that was directly related to the contemporary barley breeding genepool. We explored population structure in a large set of germplasm then used phylogeny, principle coordinates and STRUCTURE analyses to explore stratification and admixture in the germplasm, then chose to remove outlying lines from the final panel that we now use routinely for association mapping studies.

10.5 Genetic Markers

Given the increased resolution in association mapping panels to maximise the chances of exploiting it effectively, it is important that the number of molecular markers used for analysis is sufficient to exploit the number of recombination events. An early attempt at an association analysis in barley was by Kraakman and colleagues (2004). Using sparse genome coverage they reported a number of significant associations for yield and stability of yield with a number of AFLP loci. They claimed some correspondence of the position of these loci with known QTL from biparental mapping studies but this assertion was complicated by a lack of common markers. In a subsequent study using the same material they reported marker loci significantly associated with Barley Yellow Dwarf Virus resistance and quantitative measures of leaf rust resistance (Kraakman et al. 2006). Again some correspondence of positions with previous studies was claimed but in one instance the particular AFLP locus had been previously reported to be the peak marker for Rphq2, a major QTL for partial resistance to *P. hordei*. The most important limitation in these early studies was that the marker technology employed, AFLP, is not well suited to this application.

A breakthrough came with the development of highly parallel SNP assay systems such as the Illumina GoldenGate™ assay implemented with their oligo pool array technology (Fan et al. 2003; Rostoks et al. (2006) and Close et al. (2009)) used alignments between barley EST sequences to identify SNPs and used these to generate

two 1536 SNP barley oligo pool assays (BOPA1 and BOPA2). Using BOPA1 on a relatively small population of barley cultivars Rostoks et al. (2006) successfully identified associations between a cluster of CBF genes responsible for winter hardiness in barley by GWAS after classifying the genotypes according to their spring or winter growth habit. Since then, more dense arrays of markers have been produced for application in GWAS. For example, we recently exploited Illumina GAIIX RNA-seq datasets from a range of barley cultivars to identify > 30,000 robust SNPs and incorporated approximately 8,000 of these on a higher density SNP platform called a 9K iSELECT Infinium array (our unpublished results). It is likely that similar but higher density chips with > 30,000 SNPs will be developed in the near future.

However there is some debate over whether this platform is the best in the longer term. As the cost of generating high coverage genome sequence continues to drop, we and others have turned to another approach termed Genotyping-by-Sequencing (GbS) (Elshire et al. 2011). GbS promises even deeper depth of coverage of polymorphic sequence information while avoiding the serious issue of ascertainment bias inherent in SNP chip platforms (see below). The disadvantage at the current moment in time is that the informatics pipelines required to analyse GbS datasets require custom scripts, generally written by specialists in the labs pioneering the approach. In contrast, Infinium array development is accompanied by an 'out-of-the-box' software suite from the vendor that enables simple allele calling and QC along with easy export into various analytical packages. Of course, this situation will rapidly change as more individuals adopt the GbS approach.

10.6 Ascertainment Bias

The development of multiplex assays such as the Infinium chip discussed above generally involves mining data extracted from a limited number of individuals. The utility of the SNP sets thus obtained is affected by the parameters of this discovery protocol. SNPs are generally identified in a discovery panel, which consists of a small sample of individuals from a population. As this panel represents only a subset of the individuals, only a fraction of total polymorphisms will be discovered. Consequently, when these SNPs are then genotyped on a larger sample of individuals an 'ascertainment bias' is introduced (Nielsen 2000). Because the discovery panel is small, the probability that a SNP will be identified in this panel is a function of the allele frequency. Thus, rare SNPs will go undiscovered more often than common SNPs. When a SNP platform developed this way is then used to screen a much broader set of germplasm, the introduced bias may compromise measures of relatedness and genetic diversity. This is largely because statistical measures that rely on allele frequency, such as nucleotide diversity, population genetics parameters and linkage disequilibrium will be affected, and have been observed (Nielsen 2000; Schlotterer and Harr 2002; Rosenblum and Novembre 2007; Storz and Kelly 2008). In barley BOPA1, BOPA2 and the recent 9K iSelect platform have also been selected from a limited number of barley accessions (Rostoks et al. 2005, 2006; Close et al.

2009; Waugh et al. unpublished data). These SNPs have provided extensive genome coverage and have dramatically progressed our understanding of the distribution of genetic diversity within the barley gene pool. Indeed several large scale projects have already used these platforms to identify marker-trait associations in elite cultivars (AGOUEB, <http://www.agoueb.org>; BarleyCAP, <http://barleycap.cfans.umn.edu>; ExBarDiv: http://pgrc.ipk-gatersleben.de/barley/net/projects_exbardiv.php) (Waugh et al. 2010). We should be mindful that the extent and patterns of diversity observed will be limited by such ascertainment issues present in the underlying data.

Particularly problematic is the use of SNPs ascertained from the cultivated gene pool to examine diversity outside of that genetically narrow set. In barley we are fortunate to have extensive collections of wild progenitors collected from the Mediterranean basin through south western Asia and eastwards as far as Tajikistan and the Himalayas, as well as locally cultivated landraces grown throughout the marginal regions of the Fertile Crescent. Understanding the genetic diversity within these, particularly the landrace collections that grow and yield under extreme conditions of temperature and water availability, will be important in future breeding programmes that seek to respond to a range of environmental challenges.

Moragues et al. (2010) evaluated the effects of SNP number and selection strategy on estimates of germplasm diversity and population structure for different types of barley collections. Using the 1536 BOPA1 SNP data and various subsets of 384 and 96 SNPs that could in principle be used for affordable middle-throughput genotyping platforms, they compared diversity statistics for 161 landraces from Jordan and Syria with 171 European cultivars. Differences were observed in patterns of SNP polymorphisms as well as a lower estimate of diversity in the landraces, contradicting previous studies using SSRs (Russell et al. 2003). This bias could be at least partially nullified by selecting an appropriate subset of SNPs. All marker subsets gave qualitatively similar estimates of the population structure in both landraces and cultivars. Russell et al. (2011) described the first application of the BOPA1 SNP platform to assess the evolution of barley in a portion of the Fertile Crescent, by genotyping geographically matched landrace and wild barleys (448 accessions) from Jordan and Syria. The question of ascertainment bias skewing the landrace-wild comparison, through greater 'pruning' of rarely polymorphic markers in wild germplasm and generating an underestimate of genetic diversity, was addressed. While they were unable to exclude this possibility, their data did show higher levels of genetic variation in wild material suggesting that the relative pruning of SNPs in wild compared to landrace barley is most likely limited. Furthermore, the difference in diversity levels between landrace and wild barleys was similar to that found in previous work (Russell et al. 2004).

In this particular study they wanted to examine diversity across the genome and particularly in regions that have been identified as playing a role in domestication. If the effect of bias, introduced by choosing SNPs polymorphic in elite cultivars was likely to be problematic, the result would be a reduction of diversity in wild compared to landraces around the domestication genes; countering the objective of the study. They identified 141 cases where rolling diversity estimates were significantly different between wild and landraces, with diversity higher in wild material the vast

majority (132 cases). Many were in regions of the genome where domestication genes are found. With the possibility of ascertainment bias pushing the comparison in the other direction, this result therefore becomes doubly significant.

10.7 GWAS

The feasibility of mapping Mendelian traits that are determined by single major genes by GWAS using panels of barley cultivars was clearly demonstrated by mapping SNP polymorphisms in germplasm collections by LD to positions that corresponded exactly to locations previously assigned by biparental genetic mapping (Rostoks et al. 2006; Waugh et al. 2010). This approach has been subsequently extended to analysis of simple and more complex phenotypic traits

GWAS for Simple Phenotypes In the first reported study, Kraakman et al. (2006) used a Pearson correlation coefficient between vectors of the phenotypic response and genetic markers, correcting for multiple testing and population structure, to identify a significant association between the DUS character ‘rachilla hair length’ and the microsatellite BMAG223. Subsequently, we used GWAS to investigate the morphological differences that are used for the characterisation of cultivars in tests of Distinctness, Uniformity and Stability (DUS). DUS characters form a ready source of highly heritable traits that are presumed to be under the control of a limited number of major genes. Cockram et al. (2010) used 490 cultivars (both winter and spring) that had been genotyped with BOPA1 revealing 1,111 sufficiently informative markers. GWAS using a mixed model to correct for population substructure identified fifteen traits that had clearly significant associations with specific genomic regions. The majority of these traits appeared to identify a single genetic locus. They included ‘seasonal growth habit’ (1H), ‘grain lateral nerve spiculation’ (2H), ‘grain aleurone colour’ (4H), ‘hairiness of leaf sheath’ (4H), ‘rachilla hair type’ (5H), ‘ear attitude’ (5H) and ‘grain ventral furrow hair’ (6H). The positions of several of these genetic positions coincided with the previously known locations for these morphological characters, others such as the 1H position shown for seasonal growth habit were unexpected. Of particular interest was a region on chromosome 2H that was found strongly associated with a number of anthocyanin based DUS characters. They noted that the Mendelian locus *ANTHOCYANINLESS 2* (*ANT2*) had been previously reported on chromosome 2HL based on studies involving biparental crosses. Similar mapping work, with a biparental population also genotyped with BOPA1 indicated that the map location of *ANT2* coincided with the position identified in the association panel. Then they derived a composite phenotype with two character states: absence of anthocyanin coloration in all recorded tissues (awns, auricles and lemma nerves), or presence in one or more of these structures. GWAS of the composite phenotype (absence of anthocyanin coloration in all recorded tissues or presence in one or more of these structures) found the genetic interval controlling this trait to lie between 93.5 and 103.7 cM on chromosome 2H, with the peak association ($-\log_{10} p = 51.7$, marker 11_21175) at 96.8 cM.

Additional genetic markers were developed using co-linearity with rice chromosome 4 and Brachypodium (*B. distachyon*) chromosome 5, ultimately defining the ANT2 locus to within a 0.57 cM interval flanked the barley homologues of *LOC_Os04g47110* and *LOC_Os04g47020*. These flanking markers were used to identify a minimum tiling path of BACs across the interval that were then sequenced. The 260 kb interval contained eleven genes, of which eight were located at collinear positions in one or more related cereal genomes. Three gene models were identified between the flanking markers, including a strong candidate gene that showed high homology to genes at the *R/B* loci that encode proteins containing a bHLH DNA-binding domain, that have previously found to control anthocyanin pigmentation in maize.

Sequencing a 4.6 kb interval across the candidate gene *HvbHLH1* in a subset of 90 cultivars identified 69 polymorphisms arranged in 4 haplotypes, with haplotype 1 exclusive to 'white' varieties, while haplotypes 2-4 were associated with anthocyanin coloration in one or more tissues. The identified polymorphisms between the haplotype groups included eight synonymous and four non-synonymous variants, as well as a 16 bp deletion within exon 6 that results in truncation of the predicted protein upstream of the bHLH domain. Subsequent genotyping in the complete association panel established that the 16 bp deletion occurred in all cultivars lacking anthocyanin pigmentation, and not in cultivars in which anthocyanin is expressed in one or more tissues. Thus, GWAS for this Mendelian trait identified a region of the genome that with additional marker development could be reduced to only three genes, including a strong candidate gene that showed functional variation and was diagnostic for the trait (see Cockram et al. 2010 for further details).

GWAS for Simple Traits Identifies Epistatic Interactions Cockram et al. (2008) identified two epistatic loci controlling vernalisation requirement by GWAS. The panel consisted of 429 spring and winter barley varieties and was genotyped with S-SAPs and SSRs together with markers based on gene specific amplicons. The genetics of vernalization requirement in barley is relatively well characterized being controlled predominantly by two major loci: *VRN-H1* and *VRN-H2* (von Zitzewitz et al. 2005). Spring alleles are thought to be due to deletions spanning putative *cis*-elements in *VRN-H1* intron I, or to deletions of part or all of the genomic region carrying the *VRN-H2* candidate genes. There is thus an epistatic relationship between the loci with winter barleys requiring winter alleles at both *VRN-H1* and *VRN-H2* potentially making their detection problematic in GWAS. However markers for both loci were found associated with winter habit in this panel with the use of genomic control (Cockram et al. 2008) as well as allowing for population structure in the analysis. This finding confirmed the results of previous detailed bi-parental mapping studies that had furnished the GWAS investigation with the markers targeting the functional polymorphisms at *VRN-H1* and *VRN-H2*.

The lack of genomic marker coverage hampered the study of Cockram et al. (2008). Ramsay et al. (2011) used the BOPA1 and BOPA2 platforms to elucidate the control of another epistatic interaction that aligns with population sub-structure in barley; that underlying ear-row number. Barley possesses three single-flowered

spikelets at each rachis node with the alternating triplets appearing opposite each other in two ranks thus forming six files of spikelets. When all three are fertile the ear has six rows of grains but if the two outer lateral spikelets are sterile then the ear is two-rowed. The presence of six rows is controlled principally by the cloned gene *VRS1*, on chromosome 2H (Komatsuda et al. 2007) that has been known for some time to be modified by the action of *INT-C* on chromosome 4H. In germplasm surveys, the *vrs1.a* allele in six-rowed barley cultivars is generally complemented by the *Int-c.a* allele and in two-rowed cultivars *Vrs1.b* is always complemented by *int-c.b*. The presence of *int-c.b* in six-rowed cultivars (i.e. *vrs1.a*, *int-c.b*) results in the development of smaller lateral spikelets (Lundqvist et al. 1997). In normal two-rowed (i.e. *Vrs1.b*) barley, *int-c.b* suppresses anther development in the lateral spikelets. In contrast, *Int-c.a* in two-rowed cultivars (i.e. *Vrs1.b*, *Int-c.a*) causes enlarged, partially male fertile, lateral spikelets.

Row type is indicative of a major population division in barley germplasm, though some cross breeding has occurred, in particular in the development of European winter-sown barleys. Despite this population stratification, association tests of row type in 190 barley cultivars with 2473 bi-allelic genome-wide SNPs revealed associations on chromosomes 1HL, 2HL and 4HS. The association of a SNP in a gene estimated to be 0.05 cM (seven genes) distal to *VRS1* indicated that the peak on 2HL was caused by *VRS1*. This was confirmed by re-sequencing *VRS1* across the mapping panel, finding complete association with causal *vrs1.a* alleles. Direct evidence for the correspondence between the association on 4HS with *INT-C* was again complicated by a lack of common markers with previous mapping studies and the inherent difficulty in phenotyping the environmentally sensitive *intermedium* trait in bi-parental populations (Lundqvist et al. 1997). Using rice gene content and order as a proxy, further characterization of the region was once again achieved by re-sequencing PCR amplicons derived from barley orthologues of the neighboring rice genes across the association panel. This showed that a significant level of association was maintained over a region of some twenty genes that included several strong candidate genes for *INT-C*, notably the barley orthologue of maize *TEOSINTE BRANCHED 1* (*ZmTB1*). *ZmTB1* is a domestication gene and member of the TCP gene family that encodes putative basic helix-loop-helix DNA-binding proteins and whose members are involved in the control of organ growth. Resequencing confirmed that *HvTB1* contained the most significantly associated SNP and genetic mapping that placed it in the expected location. Definitive evidence that *HvTB1* was *INT-C* was obtained by re-sequencing *HvTB1* in a collection of 17 known *INT-C* mutants in a *Vrs1.b* (two-row) background. The GWAS approach thus enabled dissection of the epistatic control of row-type and high resolution mapping, and ultimately cloning of the interacting genes.

GWAS for Quantitative Traits The use of GWAS to dissect the genetic control of quantitative traits is more complex than its use for simpler traits controlled by a limited number of major genes. There are evident limitations to the power of a GWAS to determine the loci underlying a quantitative trait depending on the size and nature of the panel used as well as the complexity of the genetic control of the trait. Simulations can give some guidance to the expected limitations of the

power of a particular study (Cockram et al. 2010) as well as to the appropriateness of methodologies to allow for population structure. However, the use of a much higher density of markers and the direct relationships established between association and bi-parental studies revealed by sharing same genotyping platform have made such comparisons easier in recent studies. The functional validation of candidate genes underlying quantitative variation is more complicated than those under the control of monogenic or oligogenic traits where developmental or morphological consequences of functional genetic variation may already have been characterised through the use of mutant plant resources. Usually the knowledge of the genetic architecture of the trait in the germplasm under study is scarce, there is no reference in bi-parental populations and even when positional correspondence between bi- and multi-parent populations is observed, it is generally difficult to prove that they share the same underlying genetic determinants. The nature of the trait may hinder exploration and using rice or *Brachypodium* gene content and order as a proxy is difficult because the type of gene responsible for the trait is maybe unknown. The most robust associations for entering the validation pipeline can be prioritised by identification of the same associations in independent germplasm. Figure 10.2 shows how a significant height QTL on chromosome 3H detected in a spring barley association panel consisting of 650 lines with *de novo* height data is cross-validated in an independent dataset consisting of 230 spring lines using 15 years of historical data. The association on chromosome 3H is almost certainly due to the green revolution gene *sdw1* (Jia et al. 2009) and is co-located with the *sdw1* phenotype mapped in a mapping populations (Thomas et al. unpublished data; Malosetti et al. 2011), but other associations observed have not yet been characterised. Given the difficulties associated with validating associations with components of complex traits it is not surprising that there is little in the literature yet describing successes in this domain. However the authors are aware of several studies where components of complex traits have been resolved to gene level and validated using mutant resources (Jordi Comadran and colleagues—unpublished results).

10.8 Future Prospects

Over the past several years we, and others, have successfully assembled the molecular tools, tested various analytical approaches and ‘tuned’ our choice of biological resources to effectively take advantage of genome wide association scans. Ultimately we chose to focus on exploiting variation in the relatively narrow 2-row spring barley genepool to take advantage of the limited population substructure, to reduce the number of segregating alleles at each locus, to facilitate generation of an efficient unbiased genotyping platform and to focus on contemporary germplasm that is still exploited for breeding in the public and private sectors. This latter choice in particular has allowed us to interact effectively with those involved in crop improvement and allowed easy transfer of resources and technologies into a domain that has real impact on determining the varieties that are grown in farmers’ fields. These choices

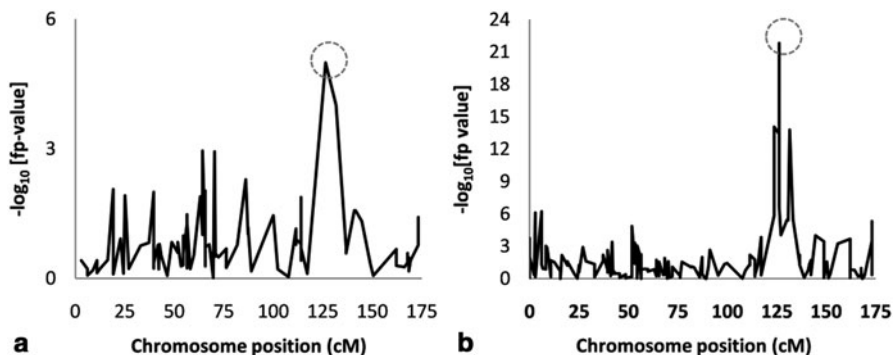


Fig. 10.2 Cross-validation of genome wide association (GWA) scans across independent germplasm sets genotyped with the same SNP platform. BOPA1 SNP loci with minimum allele frequencies > 10 % and missing data < 10 % were used for a GWAS using a kinship mixed model approach as implemented in Genstat v.14 (VSN International). TASSEL V3.0 was used to estimate the kinship matrix (K) from a subset of random markers covering the whole genome so that we did not over-estimate sub-population divergence. (a) Highly replicated height data collected from 200 elite 2 row spring cultivars over a period of ~20 years were analysed by GWAS. Several significant association peaks were detected but only chromosome 3H is shown. $-\log_{10}$ [fp values] are plotted following chromosomal order and may not reflect genetic distances. (b) Chromosome 3H scan for “*de novo*” height data collected on 650 2 row spring cultivars in one season. The top SNP (highlighted in the graphs with a circle) is tightly linked to barley green revolution gene *sdw1* (Ramsay et al. unpublished data)

together have allowed the isolation of major genes and genes controlling more complex traits. In future a significant issue remains over how we most effectively validate associations with components of highly complex traits such as yield and quality, and in such cases how the data is best exploited by the end user community. Thus, while as academics we are focused on using the information for gene identification and validation, we are also actively exploring how the phenotypic and molecular marker data can be integrated into a practical crop improvement program. Currently we are focusing on ‘Genomic Selection’ (GS—Meuwissen et al. 2001). A general view is that GS holds much promise for crop improvement but precisely how it will be implemented remains to be established. We conclude that, if establishing GWAS in barley effectively delivers the dual outcomes of facilitating gene isolation and providing the molecular and phenotypic datasets to establish Genomic Selection, then what we have learned will have been valuable and worthwhile.

References

- Aulchenko YS, de Koning D-J, de, Haley C (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177:577–585
- Caldwell KS, Russell J, Langridge P, Powell W (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *hordeum vulgare*. *Genet* 172:557–567

- Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Curr Opin Plant Biol* 11:215–221
- Close TJ, Bhat PR, Lonardi S, Wu Y, Rostoks N, Ramsay L, Druka A, Stein N, Svensson JT, Wanamaker S, Bozdogan S, Roose ML, Moscou MJ, Varshney R, Chao S, Szücs P, Sato K, Hayes PM, Matthews DE, Marshall DF, Muehlbauer GJ, Graner A, DeYoung J, Madishetty K, Fenton RD, Condamine P, Waugh R (2009) Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* 10:582
- Cockram J, White J, Leigh FJ, Lea VJ, Chiapparino E, Laurie DA, Mackay IJ, Powell W, O'Sullivan DM (2008) Association mapping of partitioning loci in barley. *BMC Genet* 9:16
- Cockram J, White J, Zuluaga D, Smith D, Comadran J, Macaulay M, Luo ZW, Kearsey MJ, Werner P, Harrap D, Tapsell C, Liu H, Hedley PE, Stein N, Schulte D, Steuernagel B, Marshall DF, Thomas WTB, Ramsay L, Mackay I, Balding DJ, Waugh R, O'Sullivan D (2010) Genome-wide association mapping of morphological traits to candidate gene resolution in the un-sequenced barley genome. *Proc Natl Acad Sci USA* 107:21611–21616
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell Sharon E (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species *PLOS ONE* 6:e19379
- Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelplnagel J, Chee MS (2003) Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* 68:69–78 (vol LXVIII)
- Feuillet C, Langridge P, Waugh R (2008) Cereal breeding takes a walk on the wild side. *Trends Genet* 24:24–32
- Jia Q, Zhang J, Westcott S, Zhang XQ, Bellgard M, Lance R, Li C (2009) GA-20 oxidase as a candidate for the semidwarf gene *sdw1/denso* in barley. *Funct Integr Genomics* 9:255–262
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB, Sabatti C, Eskin E (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–U110
- Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genet* 168:435–446
- Kraakman ATW (2005) Mapping of yield, yield stability, yield adaptability and other traits in barley using linkage disequilibrium mapping and linkage analysis. PhD dissertation 3772. Wageningen University
- Kraakman ATW, Martí'nez F, Mussiraliyev B, van Eeuwijk FA, Niks RE (2006) Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. *Mol Breed* 17:41–58
- Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, Lundqvist U, Fujimura T, Matsuoka M, Matsumoto T, Yano M (2007) Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc Natl Acad Sci USA* 104:1424–1429
- Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* 8:833–U94
- Lundqvist U, Franckowiak JD, Konishi T (1997) New and revised descriptions of barley genes. *Barley Genet Newsl* 26:22–516
- Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci* 12:57–63
- Malosetti M, van Eeuwijk FA, Boer MP, Casas AM, Elia M, Moralejo M, Ramsay L, Molina-Cano JL (2011) Gene and QTL detection in a three-way barley cross under selection by a mixed model with kinship information using SNPs. *Theor Appl Genet* 122:1605–1616
- McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, Flint-Garcia S, Thornsberry J, Acharya C, Bottoms C, Brown P, Browne C, Eller M, Guill K, Harjes K, Kroon D, Lepak

- N, Mitchell SE, Peterson B, Pressoir G, Romero S, Rosas OM, Salvo S, Yates H, Hanson M, Jones E, Smith S, Glaubitz JC, Goodman M, Ware D, Holland JB, Buckler ES (2009) "Genetic properties of the maize nested association mapping population". *Science* 325(737):737–740
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genet* 157:1819–1829
- Moragues M, Comadran J, Waugh R, Milne I, Flavell AJ, Russell JR (2010) Effects of ascertainment bias and marker number on estimations of barley diversity from high throughput SNP genotype data. *Theoretical And Applied Genetics* 120:1525–1534
- Morrell PL, Toleno DM, Lundy KE, Clegg MT (2005) Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. *Proc Natl Acad Sci USA* 102:2442–2447
- Nielsen R (2000) Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* 154:931–942
- Ramsay L, Comadran J, Druka A, Marshall DF, Thomas WTB, Macaulay M, MacKenzie K, Simpson CG, Fuller J, Bonar N, Hayes PM, Lundqvist U, Franckowiak JD, Close TJ, Muehlbauer G, Waugh R (2011) Intermedium-C, a modifier of lateral spikelet fertility in barley is an ortholog of the maize domestication gene *teosinte branched 1*. *Nature Genetic* 43:169–172
- Rosenblum EB, Novembre J (2007) Ascertainment bias in spatially structured populations: a case study in the eastern fence lizard. *J Hered* 98:331–336
- Rostoks N, Mudie S, Cardle L, Russell JR, Ramsay L, Booth A, Svensson JT, Wanamaker SI, Walia H, Rodriguez EM, Hedley PE, Liu H, Morris J, Close TJ, Marshall DF, Waugh R (2005) Genome wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Mol Genet Genomics* 274:515–527
- Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole genome association mapping in elite crop varieties. *Proc Natl Acad Sci USA* 103:18656–18661
- Russell JR, Booth A, Fuller JD, Baum M, Ceccarelli S, Grando S, Powell W (2003) Patterns of polymorphism detected in the chloroplast and nuclear genomes of barley landraces sampled from Syria and Jordan. *Theor Appl Genet* 107:413–421
- Russell J, Booth A, Fuller F, Harrower B, Hedley P, Machray G, Powell W (2004) A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the VSN International, Hemel Hempstead barley genome. *Genome* 47:389–398
- Russell JR, Dawson IK, Flavell AJ, Steffenson B, Weltzien E, Booth A, Ceccarelli S, Grando S, Waugh R (2011) Analysis of more than 1,000 SNPs in geographically-matched samples of landrace and wild barley indicates secondary contact and chromosome-level differences in diversity around domestication genes. *New Phytol* 191:564–578
- Schlotterer C, Harr B (2002) Single nucleotide polymorphisms derived from ancestral populations show no evidence for biased diversity estimates in *Drosophila melanogaster*. *Molecular Ecology* 11:947–950
- Schmalenbach I, Léon J, Pillen K (2009) Identification and verification of QTLs for agronomic traits using wild barley introgression lines. *Theor Appl Genet* 118:483–497
- Storz JF, Kelly JK (2008) Effects of spatially varying selection on nucleotide diversity and linkage disequilibrium: insights from deer mouse globin genes. *Genet* 180:367–379
- Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Sci* 277:1063–1066
- Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 28:286–289
- Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Sci* 310:1031–1034
- VSN International (2011) *GenStat for Windows*, 14th edn. VSN International, Hemel Hempstead, UK. Web page: GenStat.co.uk

- von Korff M, Wang H, Léon J, Pillen K (2005) AB-QTL analysis in spring barley. I. Detection of resistance genes against powdery mildew, leaf rust and scald introgressed from wild barley. *Theor Appl Genet* 111:583–590
- von Zitzewitz J, Szücs P, Dubcovsky J, Yan L, Francia E, Pecchioni N, Casas A, Chen THH, Hayes P, Skinner J (2005) Molecular and structural characterization of barley vernalization genes. *Plant Mol Biol* 59:449–467
- Waugh R, Marshall D, Thomas WTB, Comadran J, Russell JR, Close T, Stein N, Hayes P, Muehlbauer G, Cockram J, O'Sullivan D, Mackay I, Flavell AJ, Agoueb, BarleyCAP, Ramsay L (2010) Whole-genome association mapping in elite inbred crop varieties. *Genome* 53:967–972
- Yan L, Fu D, Li C, Blechl A, Tranquilli G, Bonafede M, Sanchez A, Valarik M, Yasuda S, Dubcovsky J (2006) The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc Natl Acad Sci USA* 103:19581–19586
- Yu J, Pressoir G, Briggs WH, Vroh I, Bi M, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordoñas JM, Buckler ES (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* 42:355–U118

Chapter 11

Production and Molecular Cytogenetic Identification of Wheat-Alien Hybrids and Introgression Lines

Márta Molnár-Láng, István Molnár, Éva Szakács, Gabriella Linc and Zoltán Bedő

Contents

11.1 Introduction	256
11.1.1 Interspecific and Intergeneric Hybridization of Plant Species.....	256
11.1.2 Molecular Cytogenetic Techniques	256
11.2 Wide Hybridization of Wheat	258
11.2.1 Wheat × Barley Hybridization	258
11.2.2 Wheat × Rye Hybrids	262
11.2.3 Wheat × <i>Aegilops</i> Hybrids	266
11.2.4 Wheat × <i>Thinopyrum</i> (syn. <i>Agropyron</i>) Hybrids	270
11.3 Conclusions	273
References	273

Abstract Barley, rye, *Aegilops* and *Thinopyrum* (syn. *Agropyron*) species belonging to the *Triticeae* tribe have large genetic diversity and serve as a valuable genetic reservoir for wheat improvement. Many of these species have been used for more than a century for the production of wheat × alien hybrids and introgression lines. The most up-to-date molecular cytogenetic techniques make it possible to detect and identify alien chromosomes in the wheat genome. The first methods used to identify rye, barley, *Aegilops* and *Thinopyrum* chromosomes in the wheat genome were C- and N-banding. Genomic *in situ* hybridization (GISH) is the most accurate way of detecting the translocation breakpoint in introgression lines. Alien chromosomes can be identified in the wheat genome using fluorescence *in situ* hybridization (FISH) with the help of repetitive DNA probes. Multicolor GISH (mcGISH) was developed to demonstrate the various genomes in polyploid plant species and in interspecific and intergeneric hybrids, amphiploids and derivatives. Sequential GISH and FISH are useful methods for identifying alien translocations in the wheat genome.

M. Molnár-Láng (✉) · I. Molnár · É. Szakács · G. Linc · Z. Bedő
Agricultural Institute, Centre for Agricultural Research, Hungarian Academy of Sciences,
P. O. Box 19, Martonvásár 2462, Hungary
e-mail: molnar.marta@agrar.mta.hu

Keywords Wheat-alien hybrids · Introgression lines · GISH · FISH · Rye · Barley · *Aegilops* · *Agropyron* · *Thinopyrum* · Triticaceae

11.1 Introduction

11.1.1 *Interspecific and Intergeneric Hybridization of Plant Species*

For several centuries scientists have been interested in hybridizing different plant species in order to merge useful traits in a new hybrid progeny. The first known artificial interspecific hybrid was produced by Fairchild in 1717 (see Belea 1992), while systematic attempts at interspecific crossing are linked with the name of Kölreuter (1766). In 1876, Stephen Wilson presented some completely sterile ears of wheat-rye hybrids for the consideration of the Botanical Society of Edinburgh. His work was aimed at the unification of the favourable characters of the two crops. This was the first step in expanding the gene pool of wheat to include the variations carried by the many wild and cultivated related species in the *Triticaceae* tribe.

The grass tribe *Triticaceae* includes some of the major cereal crop species of the world, namely *Triticum aestivum* L. (bread wheat), *T. durum* L. (durum wheat), *Secale cereale* L. (rye), *Hordeum vulgare* L. (barley), the modern cereal *Triticosecale* (triticale) and about 350 other species (Knüpffer 2009). Species related to wheat in the *Triticaceae* tribe have large genetic diversity and serve as a valuable genetic reservoir for wheat improvement. The majority of these species can be crossed with wheat and agronomic traits can be transferred from the hybrids into the wheat genome by backcrossing. In 1969, a new era of molecular cytogenetics began which made it possible to precisely identify the chromosomes of different species and to determine the genome composition of hybrids and derivatives.

11.1.2 *Molecular Cytogenetic Techniques*

11.1.2.1 Chromosome Banding Techniques

The first chromosome banding technique was developed by Caspersson et al. (1968). Alkylating fluorochromes like quinacrine (Q) and quinacrine mustard (QM) were found to differentially stain regions of C-heterochromatin in the chromosomes of *Vicia faba*. The Q and QM fluorescent patterns were chromosome-specific, thus allowing the chromosomes to be identified (see Friebe and Gill 1996). The Giemsa banding techniques originated as a by-product in an *in situ* hybridization (ISH) experiment where mouse satellite DNA was hybridized to mouse metaphase chromosomes. Pardue and Gall (1970) observed that the DNA probe preferentially hybridized to the centromeric regions, but they also observed that these regions stained darker than

other chromosome regions after counterstaining with Giemsa. This discovery led to the development of the C- and G-banding techniques for mammalian chromosomes and shortly afterwards to the development of similar techniques for plant chromosomes (Sarma and Natarajan 1973; Hadlaczký and Belea 1975). A standard karyotype and banding nomenclature system for *T. aestivum* was proposed by Gill et al. (1991). Today, it is possible to identify all 21 chromosome pairs of hexaploid wheat and also 36 of the 42 chromosome arms by C-banding. The barley chromosomes were identified by Giemsa C- and N-banding (Linde-Laursen 1975).

11.1.2.2 *In situ* Hybridization

The development of the DNA *in situ* hybridization (ISH) technique (Gall and Pardue 1969; John et al. 1969) marked the transition from the classical cytogenetics era to the modern molecular cytogenetics era (see Jiang and Gill 2006). The basic procedure of ISH is the labelling of a DNA probe and its hybridization to cytological preparations. ISH is a powerful method for localizing DNA or RNA sequences in the cytoplasm, organelles, chromosomes or nuclei of biological material (Leitch et al. 1994). Radiation-based methods were used in probe labelling and signal detection in early techniques, but they were soon replaced by fluorescence-based methodologies (Langer-Safer et al. 1982). Fluorescence *in situ* hybridization (FISH) using fluorochromes for signal detection has several advantages over ISH using isotopic probes or enzymatic detection methods. First, different DNA probes can be labelled with different haptens and simultaneously detected using different fluorochromes (multicolor FISH), thus allowing their physical order on chromosomes to be determined (Lichter et al. 1990; Leitch et al. 1991; Mukai et al. 1993). Second, fluorescence signals can be captured by special cameras or laser scanning microscopes and analysed with digital imaging systems, thus allowing more precise mapping (Jiang and Gill 1994).

Total genomic DNA probes with unlabelled blocking DNA can also be used to identify the genomes in hybrid organisms (Le et al. 1989; Schwarzachner et al. 1989). Genomic probes are used in plant breeding to detect alien translocations and substitutions in cereals (Schwarzachner et al. 1992). Genomic *in situ* hybridization (GISH) is the most efficient and accurate technique to allocate the breakpoints and estimate the amount of alien chromatin in translocation chromosomes (Anamthawat-Jonsson et al. 1993; Jiang and Gill 1994).

FISH signals derived from a single repetitive DNA probe or a cocktail containing several DNA probes can provide a hybridization pattern that allows all the chromosomes within a species to be identified. Since different probes or probe cocktails can be developed for each species, the FISH-based chromosome identification method is more versatile than the traditional chromosome banding techniques (Jiang and Gill 2006). More importantly, FISH-based chromosome identification systems can be integrated directly into the FISH mapping of other DNA sequences. Attempts to increase the detection sensitivity of very small chromosomal targets, and to improve the spatial resolution of signals derived from flanking sequences, have led to the

development of a variety of novel techniques: it is now possible to perform *in situ* hybridizations on interphase nuclei, meiotic pachytene chromosomes and isolated chromatin (DNA fibres) (de Jong et al. 1999).

11.2 Wide Hybridization of Wheat

Molecular cytogenetic techniques are applied in the selection and identification of progenies originating from distant crosses, which contain alien chromosome segments. The method for transferring genes from related species to wheat largely depends on the evolutionary distance between the species involved. Species belonging to the primary gene pool of common wheat share homologous genomes. Gene transfer from these species can be achieved by direct hybridization, homologous recombination, backcrossing and selection (Friebe et al. 1996). The secondary gene pool of common wheat includes polyploid *Triticum/Aegilops* species that have at least one homologous genome in common with *T. aestivum*. Gene transfer from these species is possible by homologous recombination if the target gene is also located on a homologous chromosome. Species belonging to the tertiary gene pool are more distantly related. Their chromosomes are not homologous to those of wheat. Other strategies need to be employed, because gene transfer from these species cannot be achieved by homologous recombination.

11.2.1 Wheat × Barley Hybridization

11.2.1.1 Production of Wheat × Barley Hybrids and Addition Lines

Bread wheat (*T. aestivum*) and barley (*H. vulgare*) are two of the most important cultivated cereals worldwide. The introgressive hybridization of barley to wheat makes it possible to transfer useful characteristics such as earliness, tolerance to drought and soil salinity, and various traits for specific nutrition quality. The first wheat × barley hybrid was produced by Kruse (1973) and the production of the first set of Chinese Spring/Betzes spring wheat-spring barley addition lines was described by Islam et al. (1978). Since then Koba et al. (1997) have reported two new 5H and 6H addition lines from a hybrid between the wheat cultivar Shinchunaga and the barley cultivar Nyugoruden. Alien additions are primarily produced to add specific desirable genes to a crop species (Gale and Miller 1987), but addition lines can be used for many other purposes, such as mapping genes and markers on introgressed alien chromosomes, examining alien gene regulation, understanding meiotic pairing behaviour and chromosome structure, and isolating individual chromosomes and genes of interest (Chang and de Jong 2005; Cho et al. 2006).

The production of wheat × barley hybrids is a difficult task because of the low crossability between the *Hordeum* and *Triticum* genera (Shepherd and Islam 1981;

Fedak and Jui 1982; Molnár-Láng and Sutka 1994). Wheat × barley hybrids can only be produced with wheat genotypes which carry recessive crossability alleles (*kr1* and *kr2*), and pollination must be carried out under favourable environmental conditions. Pollinated flowers must be given hormone treatment (2,4-dichlorophenoxyacetic acid or gibberellic acid) followed by embryo rescue, but in spite of great efforts the seed set is very low (Kruse 1973; Fedak 1980; Molnár-Láng and Sutka 1994; Molnár-Láng et al. 2000a). The hybrids are male sterile, but can be pollinated with wheat (Islam and Shepherd 1990; Koba et al. 1997). However, in most cases no backcross progenies have been obtained (Wojciechowska and Pudelska 1993; Jauhar 1995).

Fourteen winter barley and three spring barley cultivars were tested as pollinators for wheat × barley hybrid production in Martonvásár, and hybrids were successfully produced with four barley genotypes: Betzes (North American two-rowed spring barley), Igri (German, two-rowed winter barley), Osnova and Manas (Ukrainian six-rowed winter barleys). The best seed set (3.3 %) was achieved with barley cv. Betzes, but less than 1 % seed set was observed when wheat was pollinated with the barley cultivars Igri, Manas and Osnova (Molnár-Láng and Sutka 1994; Molnár-Láng et al. 2000a). There was no seed set with the other thirteen barley cultivars. The Martonvásári 9 kr1 (Mv9 kr1) × Igri and Asakaze komugi × Manas hybrids were vigorous and had good tillering ability, which made it possible to collect anthers from young inflorescences for meiotic analysis, to pollinate some spikes with wheat, and to use some developing inflorescences for *in vitro* multiplication. The hybrids were multiplied in tissue culture because of the high degree of sterility, and then pollinated with wheat to obtain backcross progenies. Meiotic analysis of the hybrids Mv9 kr1 × Igri and Asakaze × Manas and their *in vitro* regenerated progenies with the Feulgen method revealed 1.59 chromosome arm associations per cell in both initial hybrids. The number of chromosome arm associations increased after *in vitro* culture in both combinations (Molnár-Láng et al. 2000a). Wheat-barley chromosome pairing was detected in the hybrids using GISH, as in the case of wheat-rye pairing in wheat × rye hybrids (King et al. 1994; Miller et al. 1994). According to GISH analysis the number of wheat-barley chromosome arm associations increased in the *in vitro* regenerated progenies of both the wheat × barley hybrids (Molnár-Láng et al. 2000a). These results proved the possibility of producing recombinants between the two genera, and thus of transferring useful characters from barley into wheat. *In vitro* conditions caused an increase in chromosome arm association frequency in both combinations and in greater fertility in some regenerants. The Asakaze × Manas hybrids were maintained in tissue culture for several years and their meiotic pairing behaviour and genome composition were analysed after *in vitro* multiplication. The seven barley chromosomes were present in most cells, even after the third *in vitro* multiplication cycle, but some abnormalities were observed (Molnár-Láng et al. 2005).

The regenerated Mv9 kr1 × Igri hybrids were backcrossed with wheat and a series of novel winter wheat-winter barley disomic addition lines (2H, 3H, 4H, 6HS, 7H and 1HS isochromosome) were selected and identified from the selfed progenies of the BC₁ plants (Szakács and Molnár-Láng 2007, 2010a). GISH was used to confirm the presence of the barley chromosomes in the wheat genome. The barley chromosomes were identified by the FISH patterns (Fig. 11.1) obtained with various combinations

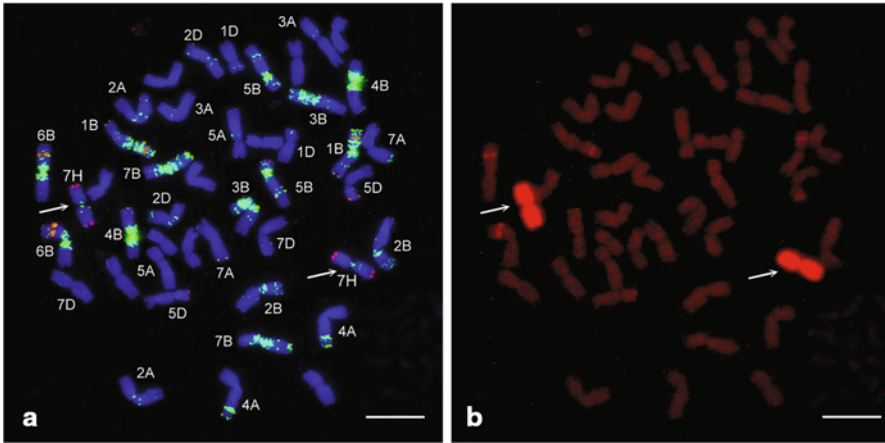


Fig. 11.1 Sequential FISH and GISH on mitotic chromosomes of 7H Mv9kr1/Igri wheat-barley disomic addition line. **a** Identification of the 7H barley chromosomes using DNA probes GAA (green), HvT01 (red) and pTa71 (yellow) on the FISH image. **b** Barley chromosomes are red as a result of labelling the barley DNA with digoxigenin and were detected with anti-DIG-Rhodamine on the GISH image. 7H barley chromosomes are indicated by arrows. Scale bar = 10 μ m

of repetitive DNA probes: GAA-HvT01, pTa71-HvT01 and Afa family-HvT01 (Szakács and Molnár-Láng 2007). Various DNA probes were used earlier to characterize the barley genome using FISH. The 45S ribosomal DNA probe pTa71 hybridizes to five chromosome pairs (Leitch and Heslop-Harrison 1992). The subtelomeric regions of all barley chromosomes can be reliably identified with the barley-specific tandem repeat HvT01 (Schubert et al. 1998) or the Triticeae-specific AT-rich tandem repeat pHvMWG2315 (Busch et al. 1995). A non-random, motif-dependent distribution of tandem array trinucleotide repeats was found for barley (Cuadrado and Jouve 2007). With the exception of (ACT)₅ the remaining trinucleotide repeats occur predominantly in Giemsa-banding-positive heterochromatin (Pedersen and Linde-Laursen 1994; Cuadrado and Jouve 2007). The identification of the barley chromosomes in the addition lines was further confirmed with SSR markers, and the addition lines were characterized morphologically.

Disomic addition lines (2H, 3H, 4H, 6H and 7H) were also selected from selfed BC₂ plants originating from the Asakaze \times Manas crosses (Molnár-Láng et al. 2012). The barley cultivar Manas is well adapted to Central European conditions, having good winter hardiness, drought tolerance and yield ability. Manas also has good tolerance of abiotic stresses such as AI and high NaCl concentration (Darkó et al. 2010; Dulai et al. 2010), so it is a suitable candidate for transferring useful agronomic traits from barley into wheat. The addition lines were identified by FISH using repetitive DNA probes (HvT01, GAA, pTa71 and Afa family), followed by confirmation with barley SSR markers. Addition lines are starting material for incorporating small segments of barley chromosomes carrying genes responsible for agronomically useful traits into the wheat genome, i.e. for producing translocation lines.

11.2.1.2 Wheat/Barley Translocations

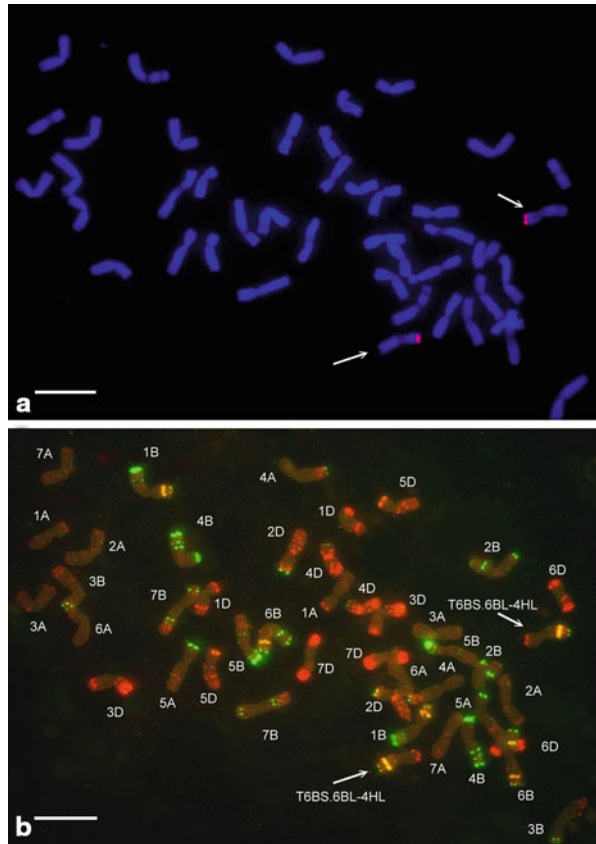
Very few wheat/barley recombinant chromosomes have been reported (Islam and Shepherd 1992), as homoeologous pairing between the chromosomes of these species is rare (Islam and Shepherd 1988; Molnár-Láng et al. 2000a). Koba et al. (1997) reported spontaneous wheat/barley translocations originating from a new wheat × barley hybrid combination. Various methods are available for producing translocations, including irradiation (Sears 1956) or genetic methods, but the most promising way is to use the 2C *Aegilops cylindrica* addition line to induce chromosome rearrangements between wheat and barley, as described by Schubert et al. (1998). This is a unique genetic system that induces frequent chromosomal structural rearrangements in common wheat by introducing gametocidal (Gc) alien chromosomes into common wheat from wild species belonging to the genus *Aegilops* (*Ae. cylindrica* and *Ae. triuncialis*) (Endo et al. 1984). The rearranged chromosomes thus induced deletions of barley chromosomes and translocations between the barley and wheat chromosomes. Lines carrying rearranged barley chromosomes are designated as ‘dissection lines’. Deletion mapping on barley chromosomes 7H, 5H and 3H was performed using barley dissection lines and barley-specific EST markers (Endo 2009). The barley dissection lines were produced from CS/Betzes addition lines, so they all carry chromosome segments from Betzes barley.

Five wheat-barley translocations (2DS.2DL-1HS, 3HS.3BL, 6BS.6BL-4HL, 4D-5HS and 7DL.7DS-5HS) were identified and characterized in Martonvásár (Molnár-Láng et al. 2000b; Nagy et al. 2002) using sequential GISH and two-colour FISH with the probes pSc119.2 and pAs1, and later by three-colour FISH with the probes pSc119.2, Afa family and pTa71 (Fig. 11.2). The barley chromatin in these lines was identified using SSR markers. The wheat/barley translocation lines were used for the physical mapping of molecular markers on barley chromosome regions (Kruppa et al. 1975).

A spontaneous interspecific Robertsonian translocation was revealed by GISH in the progenies of a monosomic 7H addition line originating from the Asakaze × Manas hybrid. FISH performed with the repetitive DNA sequences Afa family, pSc119.2 and pTa71 allowed the identification of all the wheat chromosomes, including wheat chromosome arm 4BS involved in the translocation (Cseh et al. 2011). FISH using barley telomere- and centromere-specific repetitive DNA probes (HvT01 and AGGAG) confirmed that one of the arms of barley chromosome 7H was involved in the translocation. SSR markers identified the translocated chromosome segment as 7HL. The presence of the *HvCslF6* gene, responsible for (1,3;1,4)-β-D-glucan production, was revealed in the centromeric region of 7HL. An increased (1,3;1,4)-β-D-glucan level was also detected in the translocation line, demonstrating that the *HvCslF6* gene is of potential relevance for the manipulation of wheat (1,3;1,4)-β-D-glucan levels (Cseh et al. 2011, 2013).

Addition lines are good starting material for the incorporation of small segments of barley chromosomes, carrying genes responsible for agronomically useful traits, into the wheat genome. It will be possible to produce new barley dissection lines containing chromosome segments from Igri and Manas, which may give new information for the mapping of DNA sequences related to various agronomic traits in

Fig. 11.2 Sequential GISH and FISH on mitotic chromosomes of 6BS.6BL-4HL wheat-barley disomic translocation line. **a** Detection of barley chromosome segments (*red*) in the translocation chromosome pairs using GISH. Total barley DNA was labelled with digoxigenin, and detected with anti-DIG-Rhodamine. The translocated chromosomes are indicated by arrows. The wheat chromosomes are blue as a result of counterstaining with DAPI. **b** Identification of the wheat chromosomes using FISH with DNA probes pSc119.2 (*green*), Afa family (*red*) and pTa71 (*yellow*). Disomic 6BS.6BL-4HL translocated chromosomes are indicated by arrows. Scale bar = 10 μm



barley. It will also be possible to introgress chromosome segments carrying genes for agronomically useful traits (nutritional parameters; AI, drought and salt tolerance) from the two-rowed and six-rowed winter barley cultivars Igri and Manas into wheat, and to determine the chromosomal location of these genes.

11.2.2 *Wheat* × *Rye* Hybrids

Secale is a small but important cereal genus that includes cultivated rye (*S. cereale* L.), weedy rye, and several wild species. As it is capable of producing higher yields than wheat under adverse conditions, rye has become a staple food grain at higher elevations and in regions with poor soils and severe winters. *Secale* spp. contain genes associated with resistance to many cereal diseases, winter hardiness, drought tolerance, sprouting, high lysine content and morphological and biochemical traits, which can be transferred to closely related cereal crops (Molski et al. 1985).

11.2.2.1 Wheat × Rye Crossability

The crossability of hexaploid wheat (*T. aestivum*) with rye is controlled by two loci, *Kr1* and *Kr2*, where the dominant alleles reduce crossability, *Kr1* having the more and *Kr2* the less potent effect. Plants which carry the *Kr1Kr1Kr2Kr2* dominant alleles give lower than 5 % seed set when pollinated with rye, but genotypes with the *kr1kr1kr2kr2* recessive homozygous genome composition may have over 50 % seed set with rye (Lein 1943). The *kr1* gene is located on the long arm of chromosome 5B, while *kr2* is located on the long arm of chromosome 5A (Riley and Chapman 1967; Lange and Riley 1973). Most European wheat varieties carry the dominant *Kr* alleles and thus have very low crossability with rye (Kiss and Rajháthy 1956; Zeven 1987). The recessive *kr* alleles are mostly present in wheat varieties from China, Japan, Siberia and other Asiatic regions, but these varieties are not suitable for production under Central European conditions. Snape et al. (1987) and Gay and Bernard (1994) transferred the recessive *kr1* allele into English and French varieties, respectively, by first incorporating the 5B chromosome from Chinese Spring or Fukuhokomugi into monosomic lines of these varieties. The major gene *Kr1* was identified on 5BL, and *SKr*, a strong QTL affecting crossability between wheat and rye, on chromosome 5BS (Tixier et al. 1998). Two SSR markers completely linked to *SKr* were used to evaluate a collection of crossable wheat progenies originating from primary triticale breeding programmes. The results confirm the major effect of *SKr* on crossability and the usefulness of the two markers for the efficient introgression of crossability into elite wheat varieties (Alfares et al. 2009).

In Hungary, the recessive crossability allele *kr1* was transferred from the spring wheat variety Chinese Spring (CS) into the winter wheat variety Martonvásári 9 (Mv9) by backcrossing Mv9 × CS hybrids with Mv9 (Molnár-Láng et al. 1996). As a result of five backcrosses with Mv9 and two selfings after each backcross, the selected progenies had over 50 % seed set with rye when tested on a large number of individual plants. These data confirmed that after the fifth backcross the selected Mv9 *kr1* line carried the recessive crossability alleles *Kr1* and *Kr2*, but the genotype was 98.4 % Mv9. When the Mv9 *kr1* line was pollinated with the old Hungarian rye cultivar Lovászpatonai (Molnár-Láng et al. 2002), the mean crossability percentage was fairly high, 68.4 %. The chromosome number distribution, examined in mitotic chromosome spreads of octoploid triticale obtained via colchicine treatment of the initial hybrid, was found to range from 51 to 56. All the rye chromosomes were identified with the help of C-banding and were detected using GISH (Nagy et al. 1998). The Mv9 *kr1* line is now used as a maternal partner in wheat-alien hybridization experiments in Martonvásár (Molnár-Láng et al. 2002). This has the advantage that the alien genes can be transferred directly into a winter wheat line with good yielding ability and good quality, instead of into CS, which has many unfavourable features from the agronomic point of view.

11.2.2.2 Wheat-Rye Addition and Substitution Lines

Rye is most intensively used to extend the genetic variability of wheat via intergeneric hybridization and recombination (Lelley 1993). Wheat-rye addition and substitution

lines played an important role in determining the homoeologous relationship between the two genera and were extensively used to search for useful genes in rye for wheat breeding. The first wheat-rye addition lines were produced by O'Mara (1940) and since then complete series of disomic wheat-rye addition lines, including adequate disomic telocentric lines (see Shepherd and Islam 1988; Lukaszewski 1988) have been developed. The rye chromosomes in the addition lines were detected by GISH and identified using C-banding and in situ hybridization with the help of labelled repetitive DNA probes (Mukai et al. 1992).

The genetic stability of wheat/rye disomic addition lines was checked using the Feulgen method and FISH (Szakács and Molnár-Láng 2010b). Feulgen staining detected varying proportions of disomic, monosomic and telosomic plants among the progenies. The greatest stability was observed for the 7R addition line, while the most unstable lines were those with 2R and 4R additions. Chromosome rearrangements were also detected using FISH. Based on the specific hybridization patterns of repetitive DNA probes (pSc119.2 and (AAC)₅), and ribosomal DNA probes (5S and 45S), isochromosomes were identified in the progenies of the 1R and 4R addition lines. The results draw attention to the importance of using FISH for continuous cytological checks on basic genetic materials because this method reveals chromosome rearrangements not detected either with the conventional Feulgen staining technique or with molecular markers (Szakács and Molnár-Láng 2010b).

The first reports on the spontaneous wheat-rye chromosome substitution 5R(5A) were published by Kattermann (1937) and O'Mara (1947). Driscoll and Anderson (1967) reported the substitution of wheat chromosomes 3A, 3B, 3D and 1D by rye chromosome 3R. Since then many other wheat-rye substitutions have been produced and identified (Lukaszewski 1991; Schlegel 1997).

11.2.2.3 Wheat/Rye Translocations

The 1BL.1RS wheat/rye translocation is the most widespread alien translocation, detected in hundreds of wheat cultivars worldwide (Bedö et al. 1993; Rabinovich 1998, Lukaszewski 2000). Most varieties with a 1BL.1RS translocation contain the short arm of the 1R chromosome from Petkus rye (Zeller 1973; Schlegel and Korzun 1997). Unfortunately most of the resistance genes (*Lr26*, *Yr9*, *Pm* and *Sr31*) located on this chromosome arm are no longer effective against new biotypes of the diseases. However, the translocation was also postulated to have a yield-enhancing effect and to improve adaptability (Rajaram et al. 1990; Villareal et al. 1998). As it is probably of single origin (Schlegel and Korzun 1997) this 1RS arm lacks any genetic variation, so new allelic variation needs to be introduced from other 1RS chromosomes in order to exploit the rich gene reservoir of diploid rye. Other rye genotypes may have new resistance genes or alleles against various diseases and may have a less deleterious effect on bread-making quality, probably the only negative consequence of the presence of the original Petkus rye chromosome arm in wheat.

Several authors have reported the production of wheat cultivars carrying 1RS chromosome arms from various rye genotypes. The 1RS.1AL translocation in wheat

cultivar Amigo carries the 1RS arm of Insave rye (Zeller and Fuchs 1983). Salmon, another 1BL.1RS wheat/rye translocation line, was derived from an F₃ seed from a hybrid between two octoploid triticale strains (Tsunewaki 1964). A 1DL.1RS translocation was derived from the rye cultivar Imperial (Shepherd 1973). Marais et al. (1994) used homologous recombination to transfer a gene from the short arm of chromosome 1R from Turkey 77 rye into the 1RS arm of the translocated chromosome in the wheat cultivar Veery. A new 1BL.1RS wheat/rye translocation line was developed by Ko et al. (2002) from the backcross of the F₁ hybrid of wheat cv. Olmil and rye cv. Paldanghomil. A fast, efficient method is urgently needed to introduce a substantial amount of allelic variation into this chromosome arm directly from diploid rye (Nagy et al. 2003; Lelley et al. 2004). Nagy et al. (2003) demonstrated that new genetic variation from the 1RS arm of rye can routinely be introduced into the 1RS of translocation wheats by crossing commercial cultivars, containing the 1BL.1RS chromosome, with octoploid triticale lines.

Molnár-Láng et al. (2010) developed a wheat genotype containing both the recessive crossability alleles (*kr1kr1kr2kr2*), allowing high crossability between 6 × wheat and diploid rye, and the 1BL.1RS wheat/rye translocation chromosome. This wheat genotype was used as a recipient partner in wheat × rye crosses for the efficient introduction of new allelic variation into 1RS in translocation wheats. These wheat lines were selected after crossing the wheat cultivars Mv Magdaléna and Mv Béres, carrying the 1BL.1RS translocation, with the wheat line Mv9 kr1, which carries the *kr1kr1kr2kr2* alleles. The wheat × rye F₁ hybrids produced with new recipient wheat lines involving the rye cultivar Kriszta were analysed in meiosis using GISH. Chromosome pairing between the 1BL.1RS translocation and the 1R chromosome of the rye cultivar was detected in 62.4 % of the pollen mother cells of the wheat × rye hybrids. The use of FISH with repetitive DNA probes pSc119.2, Afa family and pTa71 allowed the 1R and 1BL.1RS chromosomes to be identified (Fig. 11.3). The presence of the 1RS arm from Kriszta as well as that of Petkus was demonstrated in the F₁ hybrids using the rye SSR markers RMS13 and SCM9. Based on GISH and SSR marker analysis it was concluded that recombination had occurred between the 1RS chromosome arms of Petkus and Kriszta in the translocated chromosome in four of the 22 plants analysed. New primary 1BL.1RS translocation lines were also created with three Chinese local rye varieties (Ren et al. 2011).

The first experimental wheat/rye translocation (4B-2R) was produced in 1967 (Driscoll and Anderson 1967), but the introgression of rye genetic information into wheat most famously occurred through a spontaneous 1RS.1BL wheat/rye translocation (Mettin et al. 1973; Zeller 1973). Another wheat/rye translocation with importance for breeding was found in the Danish variety Viking, which carries a 4B-5R interchange (Schlegel et al. 1993) causing high iron, copper and zinc efficiency compared to common wheat (Schlegel 2006). The old Portuguese wheat landrace, Barbela contains small, spontaneously occurring rye segments on the long arm of 2D. This landrace shows good productivity under the low fertility conditions often associated with acid soils (Ribeiro-Carvalho et al. 1997). A large number of wheat/rye translocations were detected among the progenies of triticale × wheat crosses (Lukaszewski and Gustafson 1983), involving all seven rye chromosomes.

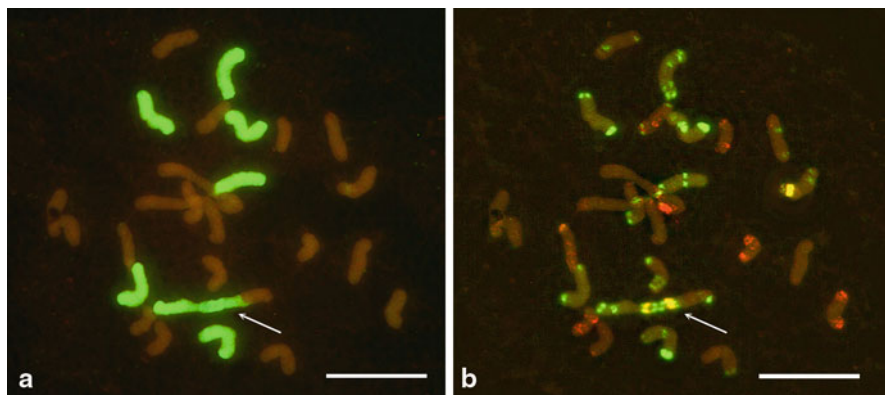


Fig. 11.3 Sequential GISH and FISH on meiotic chromosomes of a wheat \times rye hybrid. **a** Genomic *in situ* hybridization (GISH) on meiotic metaphase I chromosomes of the wheat \times rye F₁ hybrid produced between the Mv Béres kr1 wheat line having the 1BL.1RS translocation and the rye cv. Kriszta. Seven rye chromosomes and the 1RS arm in the 1BL.1RS translocated chromosome (*arrow*) are yellowish green. Twenty wheat chromosomes (18 univalents and 1 rod bivalent) and the 1BL arm of the 1BL.1RS chromosome are unlabelled. **b** 1BL.1RS-1R pairing was identified using FISH with repetitive DNA probes pSc119.2 (*green*), Afa family (*red*) and pTa71 (*yellow*) on the same cell. Scale bar = 10 μ m

Agronomically useful wheat/rye translocations were produced by incorporating chromosome segments from the 2R (Mukade et al. 1970; Sears et al. 1992; Cainong et al. 2010), 3R (Rao 1978) and 6R (Friebe and Larter 1988; Friebe et al. 1991) rye chromosomes into the wheat genome (Friebe et al. 1996).

Triticale is the first man-made crop originating from wheat \times rye hybridization (Kiss 1966; Lelley 1993). According to the FAO database it is grown on more than 4 million ha worldwide. At present, triticale is grown in Poland on 1.465 million ha, in Germany on more than 400,000 ha, in the Russian Federation on 187,000 ha and in Hungary on more than 125,000 ha.

11.2.3 Wheat \times *Aegilops* Hybrids

11.2.3.1 *Aegilops* (goatgrass) Species

The genus *Aegilops* L. comprises 11 diploid, 10 tetraploid and 2 hexaploid species (Van Slageren 1994). Some of these species took part in the evolution of pasta and bread wheat, as *Ae. tauschii* Coss. ($2n = 2 \times = 14$, DD) is the donor of the hexaploid wheat D genome and *Ae. speltoides* Tausch ($2n = 2 \times = 14$, SS) exhibits the closest relationship to the B genome of wheat (Dvorak 1998). *Aegilops* species have great diversity, thus representing a large reservoir of useful traits for wheat improvement. Species belonging to this genus have been evaluated for their resistance to diseases and pests (Gill et al. 1983, 1985; Raupp et al. 1995). Many agronomically useful

traits, including disease and pest resistance, stress and salt tolerance and winter hardiness, have been transferred from these species to wheat and several of them are used in wheat improvement (Cox et al. 1994; Gill et al. 1987; Raupp et al. 1993, see Schneider et al. 2008).

The directed exploitation of this variability requires detailed knowledge of the genetic and cytogenetic structure of the *Aegilops* species. Karyotypic data including C-banding patterns and the chromosomal distribution of four repetitive DNA sequences have been reported for all the diploid *Aegilops* species (Badaeva et al. 1996a, b). This set the stage for the analysis of the genome differentiation of the polyploid *Aegilops* species, which were analysed by C-banding and FISH with repetitive DNA probes (Badaeva et al. 2002, 2004, 2011; Schneider-Linc et al. 2005; Molnár et al. 2011).

Aegilops cylindrica Host ($2n = 4 \times = 28$, D^cD^cC^cC^c) is an autogamous, allotetraploid wild relative of bread wheat, which is native to the Mediterranean, the Middle East and Asia, and was introduced both to the Great Plains and Pacific northwest of the United States and into Hungary (Kimber and Feldman 1987; van Slageren 1994). The genomic constitution of *Ae. cylindrica* was determined by analysing chromosome pairing (Sax and Sax 1924; Kihara 1931), storage proteins (Johnson 1967), isozymes (Jaaska 1981; Nakai 1981) and differences in the restriction length patterns of repeated nucleotide sequences (Dubcovsky and Dvorak 1994). These studies identified the diploid species *Ae. caudata* L. ($2n = 2 \times = 14$, CC) as the donor of the C genome and *Ae. tauschii* as the donor of the D genome of *Ae. cylindrica*. A detailed karyotypic analysis of *Ae. cylindrica* was performed by C-banding, GISH and FISH using the DNA clones pSc119, pAs1, pTa71, and pTa794. GISH analysis detected intergenomic translocation in three of the five *Ae. cylindrica* accessions (Linc et al. 1999).

Aegilops biuncialis Vis. [syn. *Aegilops lorentii* Hochst., *T. macrochaetum* (Shuttlew. & A. Huet ex. Duval-Jouve) K. Richt] ($2n = 4 \times = 28$, U^bU^bM^bM^b) is a tetraploid wild relative of wheat belonging to the section *Polyeides* of the genus *Aegilops*. *Ae. biuncialis* shares the U and M genomes with the polyploid species *Ae. geniculata* Roth. ($2n = 4 \times = 28$, U^gU^gM^gM^g), *Ae. columnaris* Zhuk. ($2n = 4 \times = 28$, U^{co}U^{co}M^{co}M^{co}) and *Ae. neglecta* REq. Ex Bertol. ($2n = 4 \times = 28$, UⁿUⁿMⁿMⁿ). The U^b genome progenitor is the diploid *Ae. umbellulata* (syn. *Triticum umbellulatum*) Zhuk. ($2n = 2 \times = 14$, UU), while the modified M^b genome originated from *Ae. comosa* (syn. *Triticum comosum*) Sm. in Sibth. & Sm. ($2n = 2 \times = 14$, MM) (Kimber and Sears 1983; Badaeva et al. 2004). *Aegilops biuncialis* has good tolerance against biotic (Damania and Pecetti 1990; Makkouk et al. 1994) and abiotic stresses such as cold and salt stress (Colmer et al. 2006). Accessions originating from semi-desert habitats can also be used as gene sources to improve drought and heat tolerance of wheat (*T. aestivum* L.) (Molnár et al. 2004; Dulai et al. 2005). To facilitate the exact identification of the *Ae. biuncialis* chromosomes in the *T. aestivum* genetic background, FISH was carried out using repetitive DNA probes (pSc119.2, pAs1/Afa family, pTa71, (GAA)_n and (ACG)_n) on *Ae. biuncialis*, *Ae. geniculata* and their two diploid progenitor species (Schneider-Linc et al. 2005; Molnár et al. 2005, 2011a, b). Differences in the hybridization patterns (Schneider-Linc et al. 2005; Molnár et al. 2011a, b) indicated that the M genome was more variable than the U

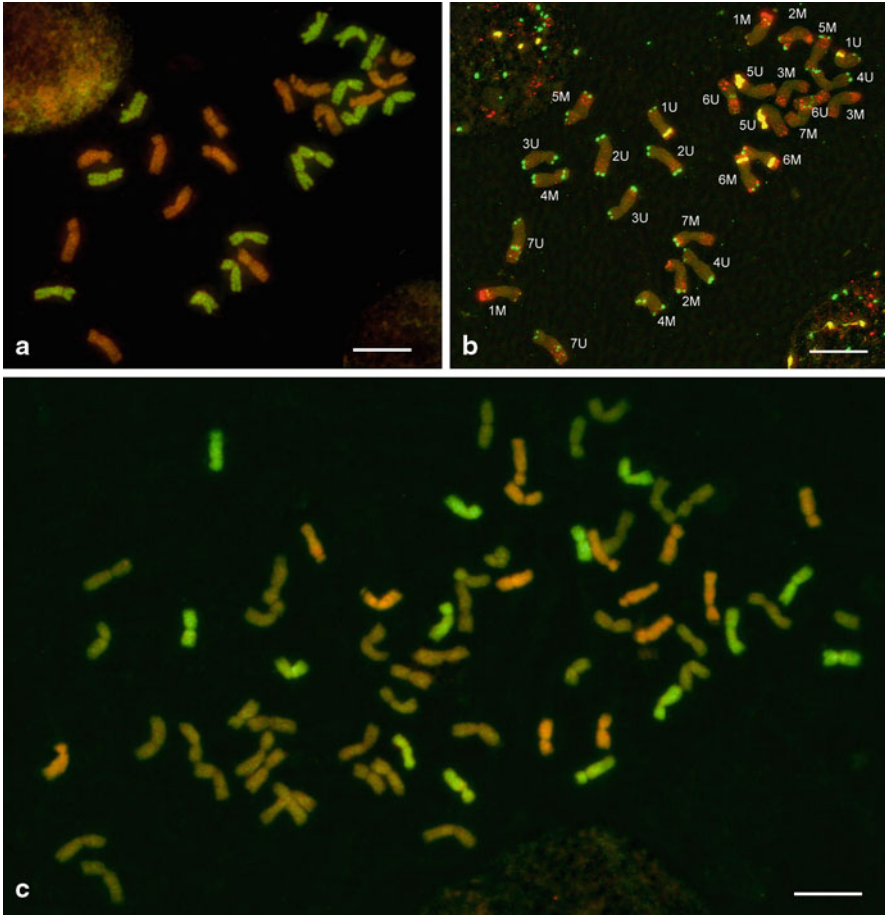


Fig. 11.4 a, b Two-colour genomic *in situ* hybridization (GISH) using U and M genomic probes, and FISH using Afa family, pSc119.2 and pTa71 repetitive DNA probes on mitotic chromosomes of *Aegilops biuncialis*. **a** On the GISH image the U genome is visualized in *orange* and the M genome in *green*. **b** On the FISH image pSc119.2 sites are *green*, Afa family signals are *red* and pTa71 signals are *yellow*. Scale bar = 10 μm . **c** Multicolor genomic *in situ* hybridization (mcGISH) on a partial wheat-*Ae. biuncialis* amphiploid cell having 39 wheat and 28 *Ae. biuncialis* chromosomes. The U^b genome is visualized in *orange*, the M^b genome in *green* and the wheat chromosomes in *brown*. Scale bar = 10 μm

genome which was confirmed by conserved orthologous set (COS) markers (Molnár et al. 2013). Intraspecific genetic variability was examined using two-colour GISH and FISH in 32 *Ae. biuncialis* (Fig. 11.4a, b) and 19 *Ae. geniculata* accessions. Homozygous intergenomic translocations were detected by GISH between the U and M genomes in six accessions (Molnár et al. 2011). Intergenomic translocation breakpoints were mapped to SSR-rich chromosomal regions (Molnár et al. 2011). The evolutionary changes in the karyotypes of the D, U and N genomes of diploid and polyploid *Aegilops* species have also been investigated by means of FISH and C-banding (Badaeva et al. 2002, 2004, 2011).

11.2.3.2 Production of Wheat × *Aegilops* Hybrids, Addition and Translocation Lines

Efforts to exploit *Aegilops* species for wheat improvement were begun more than a century ago. The results achieved to date in the field of wheat-*Aegilops* hybridization and gene transfer were reviewed by Schneider et al. (2008). Of the 23 *Aegilops* species, most of the diploids (*Ae. umbellulata* Zhuk., *Ae. mutica* Boiss., *Ae. bicornis* (Forssk.) Jaub. & Spach, *Ae. searsii* Feldman & Kislev ex Hammer, *Ae. caudata* L., *Ae. sharonensis* Eig, *Ae. speltoides* Tausch, *Ae. longissima* Schweinf. & Muschl.) and several polyploids (*Ae. ventricosa* Tausch, *Ae. peregrina* (Hack. In J. Fraser Marie & Weiller, *Ae. geniculata* Roth, *Ae. kotschyi* Boiss., *Ae. biuncialis* L.) have been used to develop wheat-*Aegilops* addition lines while wheat-*Aegilops* substitution lines have been developed using several species, including *Ae. umbellulata*, *Ae. caudata*, *Ae. tauschii*, *Ae. speltoides*, *Ae. sharonensis*, *Ae. longissima* and *Ae. geniculata* (see Kilian et al. 2011). Translocations carrying genes responsible for useful agronomic traits were developed with *Ae. umbellulata*, *Ae. comosa*, *Ae. ventricosa*, *Ae. longissima*, *Ae. speltoides* and *Ae. geniculata* (see Schneider et al. 2008).

Ae. biuncialis was crossed as male parent with the winter wheat line Mv9 kr1, and F₁ hybrids were produced with great efficiency. Amphiploids were then developed and backcrossed with wheat by Logojan and Molnár-Láng (2000) who also investigated the meiotic pairing behaviour of the hybrids. The wheat-*Ae. biuncialis* amphiploids were able to maintain significantly higher water content, photosynthetic capacity and biomass production than wheat genotypes during drought stress (Molnár et al. 2008). Six different disomic addition lines, each with 22 bivalents in metaphase I of meiosis, were selected from the selfed backcross derivatives of the amphiploids (Molnár-Láng et al. 2002). Five of them were identified using FISH with repetitive DNA probes pSc119.2 and pAs1. No chromosome rearrangements between wheat and *Ae. biuncialis* were detected by GISH in these additions (Schneider-Linc et al. 2005), which can be used to study the genetic effects of individual alien chromosomes in wheat.

Since the first successful gene transfer from *Aegilops umbellulata* Zhuk. to wheat (Sears 1956), ionising irradiation (such as X- and Y-rays) has been widely applied to crop species for the production of interspecific translocations. A large number of genes were transferred from *Aegilops* species to cultivated wheat, including those for resistance to leaf rust resistance, stem rust, yellow rust and powdery mildew, and various pests (cereal cyst nematode, root knot nematode, Hessian fly and greenbug) (see Schneider et al. 2008), but there are still many untapped genetic resources in *Aegilops* species that could be used as resistance sources for plant breeding. C-banding and FISH permit the distinction of the wheat and *Aegilops* chromosomes involved in wheat-alien translocations, whereas their size and breakpoint positions can be determined by GISH analysis (see Jiang et al. 1994; Castilho et al. 1996; see Friebe et al. 1996). Biagetti et al. (1999) used two highly repetitive DNA sequences (pSc119.2 and pAs1) and one low copy 3BS-specific RFLP sequence to physically map *Ae. longissima* chromatin in wheat recombinant lines carrying *Pm13* derived from *Ae. longissima*. Using a combination of C-banding and ISH, it was possible to

identify chromosomes carrying *Aegilops*-derived chromosome segments (see Friebe et al. 1996; Nasuda et al. 1998).

Because *Aegilops* species are more closely related to wheat than rye, barley or *Agropyron* species, it is often difficult to discriminate *Aegilops*-derived chromosomes using GISH (Wang et al. 2000; Benavente et al. 2001; Molnár et al. 2005, 2009; Cifuentes et al. 2006). However, a GISH protocol combining the pre-annealing of the probe and blocking DNA and prehybridization with blocking DNA was successfully used both to differentiate the very closely related genomes of *Ae. uniaristata* and wheat and to distinguish the S genome of *Ae. searsii* and *Ae. longissima* from the B genome of wheat (Iqbal et al. 2000; Belyayev et al. 2001). Multicolor GISH (mcGISH) using differentially labelled total genomic DNA probes enables the parental genomes to be discriminated in allopolyploid plants (Mukai et al. 1993; Belyayev et al. 2001) can also detect intergenomic chromosome rearrangements. The simultaneous visualization of individual wheat genomes and alien chromatin in interspecific hybrids and derivatives has also been reported (Sánchez-Morán et al. 1999; Han et al. 2003). Benavente et al. (2001) individually distinguished the U^o and M^o genomes of *Aegilops ovata* L. in durum wheat-*Ae. ovata* amphiploids using the total genomic DNA of *Ae. umbellulata* and *Ae. comosa* Sm. in Sibth. & Sm. as U and M genomic probes. The simultaneous discrimination of the two constituent genomes of *Ae. biuncialis* and the wheat chromosomes by mcGISH was reported by Molnár et al. (2009). This procedure also allowed for the parallel discrimination of the U^b and M^b genomes of *Ae. biuncialis* from bread wheat chromosomes (11.4). The γ -irradiation of the wheat-*Ae. biuncialis* amphiploids yielded a large number of intergenomic translocations involving the whole of the *Aegilops* and wheat genomes (Molnár et al. 2009). Dicentric chromosomes, fragments and terminal translocations were most frequently induced by γ -irradiation. Chromosome banding and ISH techniques may fail to identify translocated chromosome segments if there are no diagnostic bands or hybridization sites. In such cases chromosome-specific molecular markers may facilitate the characterization of the *Aegilops* segment.

11.2.4 *Wheat* × *Thinopyrum* (syn. *Agropyron*) Hybrids

11.2.4.1 *Agropyron* Species

Wheatgrass and wildrye grasses are some of the most important grasses in the temperate regions of the world (Wang 2011). These species are important as tertiary gene pools for wheat improvement and also serve as forage crops. Many of these grasses are related to and have been hybridized with cultivated cereal crops including wheat, barley and rye as genetic sources for disease resistance, salinity tolerance and other traits.

The taxonomy of the wheatgrass and wildrye grasses has been object of considerable controversy. The wheatgrasses traditionally have been included in the genus *Agropyron* and wildrye have been largely treated as species in the genus *Elymus*

(Wang 2011). Although it is now agreed by taxonomists that *Agropyron* should be restricted to *A. cristatum* and its close relatives, in the present review, *Agropyron* is used to include species in the genera *Australopyrum* (Tzelev) A Löve, *Dasyphyrum* (Coss. & Durieu) T. Durand, *Elymus* Linnaeus, *Leymus* Hochstetter, *Pascopyrum* A. Löve, *Pseudoroegneria* (Nevski) A. Löve, and *Thinopyrum* A Löve, etc. according to Wang (2011). All species in the genera *Agropyron*, *Pseudoroegneria*, *Psathyrostachys*, *Thinopyrum*, *Elymus* and *Leymus* are theoretically capable of being hybridized with wheat.

Species belonging to the present *Thinopyrum* genus (formerly *Agropyron*) are known to possess genes conferring resistance to various diseases, such as leaf and stem rusts, barley yellow dwarf virus, *Fusarium* head blight, etc., making these species suitable for improving the distance resistance of wheat (Friebe et al. 1994; Zhong et al. 1994; Fedak and Han 2005; Li and Wang 2009). *Thinopyrum* genus consists of three species complexes: *Th. junceum* (L.) A. Löve, *Th. elongatum* (Host) D.R. Dewey, and *Th. intermedium*. Species in this genus possess the J-, E-genome and sometimes contains the St-genome. This genus consists of diploids, segmental allotetraploids, segmental allohexaploids, octoploids and decaploids (Wang 2011).

11.2.4.2 Exploitation of *Thinopyrum* Species for Wheat Improvement

The prospect of taking advantage of the desirable gene content of these species has urged researchers worldwide to exploit the potential of the tertiary gene pool. Up till now several resistance genes have been transferred from perennial *Triticeae*, most of them originating from species in the *Thinopyrum* genus. One of the most important alien resistance genes for biotic stress transferred from *Agropyron elongatum* (syn. *Th. elongatum*) to wheat is *Lr19* (Sharma and Knott 1966). Wide hybridization of tall wheatgrass species with wheat appears promising as an avenue to improve salt tolerance (Colmer et al. 2006; Mullan et al. 2009). Other genes conferring resistance to leaf and stem rust, scab and head blight, curl mite, powdery mildew, wheat streak mosaic virus and barley yellow dwarf virus have been introgressed successfully into wheat through chromosome engineering (see Wang 2011). The two most valuable sources are *Th. intermedium* and *Th. ponticum*, firstly due to the fact that these species have resistance to rusts, common root rot, wheat scab, wheat streak mosaic virus, green bug and curl mite, and tolerance of abiotic stress (drought, high temperature, salinity) (Liu et al. 2007) and secondly because they contain the two basic genomes E and St, which are closely related to the A and D genomes of hexaploid wheat. Their polyploid nature suggests multiple origins, with progenitors from different geographical areas. They thus possess great genetic variability and molecular polymorphism which could be exploited by researchers and breeders (García et al. 2002). An intensive hybridization programme involving annual wheat and species from the former *Agropyron* genus was successfully initiated in the early 1930s by Tsitsin (Tsitsin 1960). Several *Thinopyrum*-wheat amphiploids were obtained worldwide and used for producing addition, substitution or translocation lines such as Agrotana, OK7211542, PWM706, PWMIII and PWM 209. The

GISH and multicolor GISH (mcGISH) methodologies were used to establish the cytogenetic constitution of various partial amphiploids (Chen et al. 1995, 1999; Han et al. 2004; Sepsi et al. 2008; Georgieva et al. 2011). A set of disomic addition lines was produced in each of which a chromosome for *Agropyron elongatum* (syn. *Th. elongatum*, $2n = 14$) was added to the chromosome complement of *T. aestivum* (Dvorak and Knott 1974). These were later proved to carry many agronomically useful traits (resistance to wheat streak mosaic virus, barley yellow dwarf virus, common root rot, *Fusarium* head blight, tan spot and *Stagonospora nodorum*) originating from the *Th. ponticum* progenitor and have been exploited as alien sources of disease resistance in wheat improvement (Chen et al. 1998; Thomas et al. 1998; Li et al. 2003; Fedak and Han 2005; Oliver et al. 2006).

McGISH and FISH were used to characterize the genomic composition of the wheat-*Th. ponticum* partial amphiploid BE-1. The amphiploid is a high-protein line having resistance to leaf rust and powdery mildew and has a total of 56 chromosomes per cell (Szalay 1979). Multicolor GISH identified 16 chromosomes originating from *Th. ponticum* and 14 A-genome, 14 B-genome and 12 D-genome chromosomes from wheat. Six of the *Th. ponticum* chromosomes carried segments differing from the J genome in their centromeric regions. Using the Afa family, pSc119.2 and pTa71 probes, FISH identified all the wheat chromosomes present and determined which chromosomes were involved in the translocations. On the basis of their multicolour FISH patterns, the alien chromosomes could be arranged in eight pairs and could also be unequivocally differentiated from each other (Sepsi et al. 2008). *In situ* hybridization techniques, combined with SSR marker analysis, are extremely useful in detecting and identifying intergenomic rearrangements in the wheat genome, leading to the selection of genetic materials that could be useful for future mapping studies (Somers et al. 2004; Sepsi et al. 2009).

Using the FISH technique with various repetitive DNA probes, the genes controlling agronomically important traits can be assigned to precise chromosomal regions, thus facilitating effective gene transfer. Hsiao et al. (1986) studied the karyotypes of 22 diploid species of perennial *Triticeae* representing the P, St, J (E), H, I, Ns, W and R genomes. The C-banding patterns established for 10 diploid species (Endo and Gill 1984) drew the attention to the equivalence of the J and E genomes. The two genomes are indistinguishable on the basis of the chloroplast sequence data, whereas the chromosome pairing pattern in meiosis, karyotype differences and data on the 5S DNA spacer and ITS sequences provide clear evidence that they represent different genera (Jauhar 1990; Kellogg et al. 1996). The rapid, accurate identification of these materials can only be achieved by generating detailed karyotypes of the individual genomes based on the use of molecular cytogenetic probes. A detailed FISH karyotype of the E genome of *Elytrigia elongata* (Host) Nevski (= *Agropyron elongatum*, *Thinopyrum elongatum*, $2n = 2 \times = 14$, EE) was generated and verified in several accessions using highly repetitive DNA sequences and the sequential GISH – mcFISH technique (Linc et al. 2012).

11.3 Conclusions

Alien chromatin originating from species in the tertiary gene pool can be detected in the wheat genome using GISH, as their genomes are not homologous with wheat. A GISH protocol for the detection of barley and rye chromosomes in a wheat background was elaborated more than twenty years ago and routinely applied in many laboratories. Differentiating the chromosomes of wheat and *Aegilops*, species previously classified in the same genus using GISH is a considerable challenge. The identification of wheat, barley, rye and *Aegilops* chromosomes was first achieved by C- and N-banding and later by in situ hybridization using various DNA probes (pSc119, pAs1, pTa71, pT794 HvT01, GAA, pSc250, pSc200, etc.). Sequential GISH and FISH, a combination of GISH and chromosome identification with the use of repetitive DNA probes is a very efficient method for detecting and identifying alien chromatin in the wheat genome.

Barley, rye, *Aegilops* and *Thinopyrum* (syn. *Agropyron*) species are important gene sources for wheat improvement, but have only been partially exploited. Very few reports have been published on gene transfer from barley, and most of these projects involved a single barley cultivar, Betzes. Rye is the related species exploited most frequently for wheat improvement. Several genes have been transferred from the *Aegilops* and *Agropyron* species into wheat, but there is still a vast reservoir of species, both in gene banks and in natural habitats, which could be tapped in future to enhance genetic diversity of wheat.

References

- Aghaee-Sarbarzeh M, Ferrahi M, Singh S et al (2002) *Ph¹*-induced transfer of leaf and stripe rust-resistance genes from *Aegilops triuncialis* and *Ae. geniculata* to bread wheat. *Euphytica* 127:377–382. doi:10.1023/A:1020334821122
- Alfares W, Bouguennec A, Balfourier F, Gay G, Bergès H, Vautrin S, Sourdille P, Bernard M, Feuillet C (2009) Fine mapping and marker development for the crossability gene *SKr* on chromosome 5BS of hexaploid wheat (*Triticum aestivum* L.). *Genetics* 183(2):469–481. doi:10.1534/genetics.109.107706
- Anamthawat-Jonsson K, Schwarzacher T, Heslop-Harrison JS (1993) Isolation and characterization of genome-specific DNA sequences in *Triticeae* species. *Mol Gen Genet* 240:151–158. doi:10.1007/BF00277052
- Badaeva ED, Friebe B, Gill BS (1996a) Genome differentiation in *Aegilops*. 1. Distribution of highly repetitive DNA sequences on chromosomes of diploid species. *Genome* 39:293–306. doi:10.1139/g96-040
- Badaeva ED, Friebe B, Gill BS (1996b) Genome differentiation in *Aegilops*. 2. Physical mapping of 5S and 18S-26S ribosomal RNA gene families in diploid species. *Genome* 39:1150–1158. doi:10.1139/g96-145
- Badaeva ED, Amosova AV, Muravenko OV et al (2002) Genome differentiation in *Aegilops*. 3. Evolution of the D-genome cluster. *Plant Syst Evol* 231:163–190. doi:10.1007/s006060200018
- Badaeva ED, Amosova AV, Samatadze TE et al (2004) Genome differentiation in *Aegilops*. 4. Evolution of the U-genome cluster. *Plant Syst Evol* 246:45–76. doi:10.1007/s00606-003-0072-4

- Badaeva ED, Dedkova OS, Zoshchuk SA et al (2011) Comparative analysis of the N-genome in diploid and polyploid *Aegilops* species. *Chrom Res* 19:541–548. doi:10.1007/s10577-011-9211-x
- Bedő Z, Balla L, Szunics L et al (1993) Agronomic properties of Martonvásár wheat varieties bearing the 1B/1R translocation. (A martonvásári 1B/1R transzlokációs búzafajták agronómiai tulajdonságai. Abstract in English) *Növénytermelés* 42:391–398
- Belea A (1992) Interspecific and intergeneric crosses in cultivated plants. Akadémiai Kiadó, Budapest, p 255
- Belyayev A, Raskina O, Nevo E (2001) Detection of alien chromosomes from S-genome species in the addition/substitution lines of bread wheat and visualization of A-, B- and D-genomes by GISH. *Hereditas* 135:119–122. doi:10.1111/j.1601-5223.2001.00119.x
- Benavente E, Alix K, Dusautoir JC et al (2001) Early evolution of the chromosomal structure of *Triticum turgidum*-*Aegilops ovata* amphiploids carrying and lacking the *Ph1* gene. *Theor Appl Genet* 103:1123–1128. doi:10.1007/s001220100666
- Biagetti M, Vitellozzi F, Ceoloni C (1999) Physical mapping of wheat-*Aegilops longissima* breakpoints in mildew-resistant recombinant lines using FISH with highly repeated and low-copy DNA probes. *Genome* 42:1013–1019. doi:10.1139/gen-42-5-1013
- Busch W, Martin R, Herrmann RG, Hohmann U (1995) Repeated DNA sequences isolated by microdissection. I. Karyotyping of barley (*Hordeum vulgare* L.). *Genome* 38:1082–1090. doi:10.1139/g95-144
- Cainong JC, Zavatsky LE, Chen MS et al (2010) Wheat-rye T2BS-2BL-2RL recombinants with resistance to Hessian Fly (H21). *Crop Sci* 50:920–925. doi:10.2135/cropsci2009.06.0310
- Caspersson T, Farber S, Foley GE et al (1968) Chemical differentiation along metaphase chromosomes. *Exp Cell Res* 49:219–222. doi:10.1016/0014-4827(68)90538-7
- Castilho A, Miller TE, Heslop-Harrison JS (1996) Physical mapping of translocation breakpoints in a set of wheat-*Aegilops umbellulata* recombinant lines using in situ hybridization. *Theor Appl Genet* 93:816–825. doi:10.1007/BF00224081
- Chang SB, de Jong H (2005) Production of alien chromosome additions and their utility in plant genetics. *Cytogenet Genome Res* 109:335–343. doi:10.1159/000082417
- Chen Q, Conner RL, Laroche A (1995) Identification of the parental chromosomes of the wheat-alien amphiploid Agrotana by genomic in situ hybridization. *Genome* 38:1163–1169. doi:10.1139/g95-154
- Chen Q, Conner RL, Ahmad F et al (1998) Molecular characterization of the genome composition of partial amphiploids derived from *Triticum aestivum* × *Thinopyrum ponticum* and *T. aestivum* × *Th. intermedium* as sources of resistance to wheat streak mosaic virus and its vector, *Aceria tosichella*. *Theor Appl Genet* 97:1–8. doi:10.1007/s001220050860
- Chen Q, Conner RL, Laroche A et al (1999) Genomic in situ hybridization analysis of *Thinopyrum* chromatin in a wheat-*Th. intermedium* partial amphiploid and six derived chromosome addition lines. *Genome* 42:1217–1223. doi:10.1139/gen-42-6-1217
- Cho S, Garvin DF, Muehlbauer (2006) Transcriptome analysis and physical mapping of barley genes in wheat-barley addition lines. *Genetics* 172:1277–1285. doi:10.1534/genetics.105.049908
- Cifuentes M, Blein M, Benavente E (2006) A cytomolecular approach to assess the potential of gene transfer from a crop (*Triticum turgidum* L.) to a wild relative (*Aegilops geniculata* Roth.). *Theor Appl Genet* 112:657–664. doi:10.1007/s00122-005-0168-z
- Colmer TD, Flowers TJ, Munns R (2006) Use of wild relatives to improve salt tolerance in wheat. *J Exp Bot* 57:1059–1078. doi:10.1093/jxb/erj124
- Cox TS, Raupp WJ, Gill BS (1994) Leaf rust-resistance genes *Lr41*, *Lr42*, and *Lr43* transferred from *Triticum tauschii* to common wheat. *Crop Sci* 34:339–343. doi:10.2135/cropsci1994.0011183X003400020005x
- Cseh A, Kruppa K, Molnár I et al (2011) Characterization of a new 4BS.7HL wheat/barley translocation line using GISH, FISH and SSR markers and its effect on the β -glucan content of wheat. *Genome* 54:795–804. DOI: 10.1139/g11-044

- Cseh A, Soós V, Rakszegi M, Türkösi E, Balázs E, Molnár-Láng M (2013) Expression of *HvCslF9* and *HvCslF6* barley genes in the genetic background of wheat and their influence on the wheat β -glucan content. *Ann Appl Biol* 163:142–150. doi:10.1111/aab.12043
- Cuadrado A, Jouve N (2007) The nonrandom distribution of long clusters of all possible classes of trinucleotide repeats in barley chromosomes. *Chromosome Res* 15:711–720. doi:10.1007/s10577-007-1156-8
- Damania AB, Pecetti L (1990) Variability in a collection of *Aegilops* species and evaluation for yellow rust resistance at two locations in Northern Syria. *J Genet Breed* 44:97–102
- Darkó É, Molnár-Láng M, Barnabás B (2010) Aluminium tolerance in wheat/barley introgression lines and in their parental genotypes. Society for Experimental Biology Annual Main Meeting, Abstracts, 30th June–3rd July, 2010. Prague, pp 359
- de Jong JH, Fransz P, Zabel P (1999) High resolution FISH in plants – techniques and applications. *Trends Plant Sci* 4:258–263
- Driscoll CJ, Anderson LM (1967) Cytogenetic studies in Transec—a wheat-rye translocation line. *Can J Genet Cytol* 9:375–380. doi:10.1139/g67-038
- Dubcovsky J, Dvorak J (1994) Genome origins of *Triticum cylindricum*, *Triticum triunciale*, and *Triticum ventricosum* (Poaceae) inferred from variation in restriction patterns of repeated nucleotide sequences: a methodological study. *Am J Bot* 81:1327–1335. doi:10.2307/2445408
- Dulai S, Molnár I, Prónay J et al (2005) Effects of drought on thermal stability of photosynthetic apparatus in bread wheat and in *Aegilops* species originating from various habitats. *Acta Biol Szegediensis* 49:215–217
- Dulai S, Molnár I, Haló B, Molnár-Láng M (2010) Photosynthesis in the 7H Asakaze komugi/Manas wheat/barley addition line during salt stress. *Acta Agron Hung* 58:367–376. doi:10.1556/AAgr.58.2010.4.5
- Dvorak J (1998) Genome analysis in the *Triticum*—*Aegilops* alliance. *Proc 9th Int Wheat Genet Symp*, Saskatoon, Saskatchewan, Canada pp 8–11
- Dvorák J, Knott DR (1974) Disomic and ditelosomic additions of diploid *Agropyron elongatum* chromosomes to *T. aestivum*. *Can J Genet Cytol* 16:399–417. doi:10.1139/g74-043
- Endo TR (1988) Induction of chromosomal structural changes by a chromosome of *Aegilops cylindrica* L. in common wheat. *J Hered* 79:366–370
- Endo TR (2009) Cytological dissection of barley genome by the gametocidal system. *Breed Sci* 59:481–486. doi:10.1270/jsbbs.59.481
- Endo TR, Gill BS (1984) The heterochromatin distribution and genome evolution in diploid species of *Elymus* and *Agropyron*. *Can J Genet Cytol* 26:669–678. doi:10.1139/g84-106
- Fedak G (1980) Production, morphology and meiosis of reciprocal barley-wheat hybrids. *Can J Genet Cytol* 22:117–123. doi:10.1139/g80-014
- Fedak G, Han F (2005) Characterization of derivatives from wheat-*Thinopyrum* wide crosses. *Cytogenet Genome Res* 109(1–3):360–367
- Fedak G, Jui PY (1982) Chromosomes of Chinese Spring wheat carrying genes for crossability with Betzes barley. *Can J Genet Cytol* 24:227–233. doi:10.1139/g82-024
- Friebe B, Gill BS (1996) Chromosome banding and genome analysis in diploid and cultivated polyploid wheats. In: Jauhar PP (ed) *Methods of genome analysis in plants*. CRC Press, Inc, Boca Raton, Florida, pp 39–60
- Friebe B, Larter EN (1988) Identification of a complete set of isogenic wheat-rye D-genome substitution lines by means of Giemsa C-banding. *Theor Appl Genet* 76:473–479. doi:10.1007/BF00265353
- Friebe B, Hatchett JH, Gill BS et al (1991) Transfer of Hessian fly resistance from rye to wheat via radiation induced terminal and intercalary chromosomal translocations. *Theor Appl Genet* 83:33–40. doi:10.1007/BF00229223
- Friebe B, Jiang J, Knott DR, Gill BS (1994) Compensation indices of radiation-induced wheat-*Agropyron elongatum* translocations conferring resistance to leaf rust and stem rust. *Crop Sci* 34:400–404. doi:10.2135/cropsci1994.0011183X003400020018x

- Friebe B, Jiang J, Raupp WJ et al (1996) Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. *Euphytica* 91:59–87. doi:10.1007/BF00035277
- Gale MD, Miller TE (1987) The introduction of alien genetic variation into wheat. In: Lupton FGH (ed) *Wheat Breeding: Its scientific basis*. Chapman and Hall, UK, pp 173–210
- Gall JG, Pardue ML (1969) Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc Natl Acad Sci USA* 63:378–383. doi:10.1073/pnas.63.2.378
- García P, Monte JV, Casanova C, Soler C (2002) Genetic similarities among Spanish populations of *Agropyron*, *Elymus* and *Thinopyrum*, using PCR-based markers. *Genet Resour Crop Evol* 49:103–109. doi:10.1023/A:1013898119274
- Gay G, Bernard M (1994) Production of intervarietal substitution lines with improved interspecific crossability in the wheat cv Courtot. *Agronomie* 14:27–32. doi:10.1051/agro:19940103
- Georgieva M, Sepsí A, Tyankova N, Molnár-Láng M (2011) Molecular cytogenetic characterization of two high protein wheat-*Thinopyrum intermedium* partial amphiploids. *J Appl Gen* 52:269–277. doi:10.1007/s13353-011-0037-1
- Gill BS, Friebe B (2009) Cytogenetic analysis of wheat and rye genomes. In: Feuillet C, Muehlbauer GJ (eds) *Genetics and genomics of the Triticeae*. Springer, Dordrecht, pp 121–135. doi:10.1007/978-0-387-77489-3_4
- Gill BS, Kimber G (1974) Giemsa C-banding evolution of wheat. *Proc Nat Acad Sci USA* 71:4086–4090. doi:10.1073/pnas.71.10.4086
- Gill BS, Browder E, Hatchett JH et al (1983) Disease and insect resistance in wild wheats. *Proc 6th Int Wheat Genet Symp*, Kyoto, Japan, pp 785–792
- Gill BS, Friebe B, Endo TR (1991) Standard karyotype and nomenclature system for description of chromosome bands and structural aberrations in wheat (*Triticum aestivum* L.). *Genome* 34:830–839. doi:10.1139/g91-128
- Gill BS, Hatchett JH, Raupp WJ (1987) Chromosomal mapping of Hessian fly resistance gene *HL3* in the D genome of wheat. *J Heredity* 78:97–100
- Gill BS, Sharma HC, Raupp WJ et al (1985) Evaluation of *Aegilops* species for resistance to wheat powdery mildew, wheat leaf rust, Hessian fly and greenbug. *Plant Breeding* 69:314–316. doi:10.1094/PD-69-314
- Hadlaczky G, Belea A (1975) C-banding in wheat evolutionary cytogenetics. *Plant Sci Lett* 4:85–88. doi:10.1016/0304-4211(75)90252-7
- Han FP, Fedak G, Benabdelmouna A et al (2003) Characterization of six wheat x *Thinopyrum intermedium* derivatives by GISH, RFLP, and multicolor GISH. *Genome* 46:490–495. doi:10.1139/g03-032
- Han FP, Liu B, Fedak G, Liu Z (2004) Genomic constitution and variation in five partial amphiploids of wheat-*Thinopyrum intermedium* as revealed by GISH, multicolour GISH and seed storage protein analysis. *Theor Appl Genet* 109:1070–1076. doi:10.1007/s00122-004-1720-y
- Hsiao C, Wang RRC, Dewey DR (1986) Karyotype analysis and genome relationships of 22 diploid species in the tribe *Triticeae*. *Can J Genet Cytol* 28:109–120
- Islam AKMR, Shepherd KW (1988) Induced pairing between wheat and barley chromosomes. In: Miller TE, Koebner RMD (eds) *Proc 7th Int Wheat Genet Symp*. England, Cambridge, pp 309–314
- Islam AKMR, Shepherd KW (1990) Incorporation of barley chromosomes into wheat. In: Bajaj YPS (ed) *Biotechnology in agriculture and forestry*, Vol 13. Wheat. Springer-Verlag, Berlin Heidelberg, pp 128–151
- Islam AKMR, Shepherd KW (1992) Production of wheat-barley recombinant chromosomes through induced homoeologous pairing. 1. Isolation of recombinants involving barley arms 3HL and 6HL. *Theor Appl Genet* 83:489–494. doi:10.1007/BF00226538
- Islam AKMR, Shepherd KW, Sparrow DHB (1978) Production and characterization of wheat-barley addition lines. In: Ramunujam S (ed) *Proc 5th Int Wheat Genet Symp*. India, New Delhi, pp 356–371

- Iqbal N, Reader SM, Caligari PDS, Miller TE (2000) Characterization of *Aegilops uniaristata* chromosomes by comparative DNA marker analysis and repetitive DNA sequence in situ hybridization. *Theor Appl Genet* 101:173–1179. doi:10.1007/s001220051594
- Jaaska V (1981) Aspartate aminotransferase and alcohol dehydrogenase isozymes: intraspecific differentiation in *Aegilops tauschii* and the origin of the D genome polyploids in the wheat group. *Plant Syst Evol* 137:259–273. doi:10.1007/BF00982790
- Jauhar PP (1990) Dilemma of genome relationship in the diploid species *Thinopyrum bessarabicum* and *Thinopyrum elongatum* (Triticeae:Poaceae). *Genome* 33:944–946
- Jauhar PP (1995) Morphological and cytological characteristics of some wheat × barley hybrids. *Theor Appl Genet* 90:872–877. doi:10.1007/BF00222025
- Jiang J, Gill BS (1994) Nonisotopic in situ hybridization and plant genome mapping: the first 10 years. *Genome* 37:717–725. doi:10.1139/g94-102
- Jiang J, Gill BS (2006) Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. *Genome* 49:1057–1068. doi:10.1139/G06-076
- Jiang JM, Friebe B, Gill BS (1994) Recent advances in alien gene-transfer in wheat. *Euphytica* 73:199–212. doi:10.1007/BF00036700
- John HA, Birnstiel ML, Jones KW (1969) RNA-DNA hybrids at the cytological level. *Nature* (London) 223:582–587 doi:10.1038/223582a0
- Johnson BL (1967) Confirmation of the genome donors of *Aegilops cylindrica*. *Nature* 216:859–862. doi:10.1038/216859a0
- Katterman G (1937) Zur Cytologie halmbehaarter Stämme aus Weizenroggenbastardierung. *Züchter* 9:196–199. doi:10.1007/BF01884284
- Kellogg EA, Appels R, Mason-Gamer RJ (1996) When genes tell different stories: the diploid genera of Triticeae. *Syst Bot* 21:321–347. doi:10.2307/2419662
- Kihara H (1931) Genomanalyse bei *Triticum* und *Aegilops*. II. *Aegilotricum* und *Aegilops cylindrica*. *Cytologia* 2:106–156
- Kilian B, Mammen K, Millet E et al (2011) *Aegilops*. In: Kole C (ed) Wild crop relatives: genomic and breeding resources. Cereals. Springer-Verlag, Berlin Heidelberg, pp 1–76. doi:10.1007/978-3-642-14228-4_1
- Kimber G, Feldman M (1987) Wild wheat, an introduction. Special Report 353, College of Agriculture, University of Missouri–Columbia, USA
- Kimber G, Sears ER (1983) Assignment of genome symbols in the *Triticeae*. *Proc 6th Int Wheat Genet Symp*, Kyoto, Japan, pp 1195–1196
- King IP, Reader SM, Purdie KA et al (1994) A study of the effect of a homoeologous pairing promoter on chromosome pairing in wheat/rye hybrids using genomic in situ hybridization. *Heredity* 72:318–321
- Kiss Á (1966) Neue Richtung in der Triticale-Züchtung. *Z Pflanzenzüchtg* 55:309–329
- Kiss Á, Rajháthy T (1956) Untersuchungen über die Kreuzbarkeit innerhalb des Subtribus Triticinae. *Züchter* 26:127–136. doi:10.1007/BF00713460
- Knüpfper H (2009) Triticeae genetic resources in ex situ genebank collections. In: Feuillet C, Muehlbauer GJ (eds) Genetics and genomics of the Triticeae. Springer, Dordrecht, pp 31–79. doi:10.1007/978-0-387-77489-3_2
- Kruppa K, Sepsí A, Szakács É, Röder MS, Molnár-Láng M (2013) Characterization of a 5HS-7DS.7DL wheat-barley translocation line and physical mapping of the 7D chromosome using SSR markers. *J Appl Genetics* 54:251–258. doi:10.1007/s13353-013-0152-2
- Kruse A (1973) *Hordeum* × *Triticum* hybrids. *Hereditas* 73:157–161. doi:10.1111/j.1601-5223.1973.tb01078.x
- Ko JM, Seo BB, Suh DY et al (2002) Production of a new wheat line possessing the 1BL.1RS wheat-rye translocation derived from Korean rye cultivar Paldanghomil. *Theor Appl Genet* 104:171–176. doi:10.1007/s00122-001-0783-2
- Koba T, Takumi S, Shimada T (1997) Isolation, identification and characterization of disomic and translocated barley chromosome addition lines of common wheat. *Euphytica* 96:289–296. doi:10.1023/A:1003081619338

- Kölreuter JG (1761–1766) Vorläufige Nachricht von einigen das Geschlecht der Pflanzen betreffenden Versuchen, und Beobachtungen, nebst Fortsetzungen 1, 2 und 3. In: Ostwald's Klassiker der Exacten Wissenschaften No 41. Verlag & Engelmann, Leipzig
- Lange W, Riley R (1973) The position on chromosome 5B of wheat of the locus determining crossability with rye. *Genet Res* 22:143–153. doi:10.1017/S0016672300012933
- Langer-Safer PR, Levine M, Ward DC (1982) Immunological method for mapping genes on *Drosophila* polytene chromosomes. *Proc Natl Acad Sci U S A* 79:4381–4385. doi:10.1073/pnas.79.14.4381
- Le HT, Armstrong KC, Miki B (1989) Detection of rye DNA in wheat-rye hybrids and wheat translocation stocks using total genomic DNA as a probe. *Plant Mol Biol Rep* 7:150–158. doi:10.1007/BF02669770
- Lelley T (2006) Triticale: a low-input cereal with untapped potential. In: Singh RJ, Jauhar PP (eds), Genetic resources, chromosome engineering, and crop improvement. Cereals, Volume 2. CRC Press, Taylor and Francis, Boca Raton, Florida, USA, pp 395–430. doi:10.1201/9780203489260.ch13
- Lelley T, Eder C, Grausgruber H (2004) Influence of 1BL.1RS wheat-rye chromosome translocation on genotype by environment interaction. *J Cereal Sci* 39:313–320. doi:10.1016/j.jcs.2003.11.003
- Lein A (1943) Die genetische Grundlage der Kreuzbarkeit zwischen Weizen und Roggen. *Zeitschr. induct. Abstamm. und Vererb. Lehre* 81:28–81. doi:10.1007/BF01847441
- Leitch IJ, Heslop-Harrison JS (1992) Physical mapping of the 18S–5.8S–26S rRNA genes in barley by in situ hybridization. *Genome* 35:1013–1018. doi:10.1139/g92-155
- Leitch AR, Schwarzacher T, Jackson D, Leitch IJ (1994) In situ hybridization: a practical guide. Bios Scientific Publishers, Oxford, UK, p 118
- Leitch IJ, Schwarzacher T, Mosgöller W et al (1991) Parental genomes are separated throughout the cell cycle in a plant hybrid. *Chromosoma* 101:206–213. doi:10.1007/BF00365152
- Li H, Wang X (2009) *Thinopyrum ponticum* and *Th. intermedium*: the promising source of resistance to fungal and viral diseases of wheat. *J Genet Genom Res* 36:557–565. doi:10.1016/S1673-8527(08)60147-2
- Li H, Chen Q, Conner RL et al (2003) Molecular characterization of a wheat–*Thinopyrum ponticum* partial amphiploid and its derivatives for resistance to leaf rust. *Genome* 46:906–913. doi:10.1139/g03-053
- Lichter P, Tang CC, Call K et al (1990) High resolution mapping of human chromosome 11 by in situ hybridization with cosmid clones. *Science (Washington DC)* 247:64–69. doi:10.1126/science.2294592
- Linc G, Friebe BR, Kynast RG et al (1999) Molecular cytogenetic analysis of *Aegilops cylindrica* Host. *Genome* 42:497–503. doi:10.1139/g98-151
- Linc G, Sepsi A, Molnár-Láng M (2012) A FISH karyotype to study chromosome polymorphisms for the *Elytrigia elongata* E genome. *Cytogenet Genome Res* 136:138–144
- Linde-Laursen I (1975) Giemsa C-banding of the chromosomes of 'Emir' barley. *Hereditas* 81:285–289. doi:10.1111/j.1601-5223.1975.tb01040.x
- Liu Z, Li DY, Zhang XY (2007) Genetic relationships among five basic genomes St, E, A, B and D in *Triticeae* revealed by genomic southern and in situ hybridization. *J Integr Plant Biol* 49:1080–1086. doi:10.1111/j.1672-9072.2007.00462.x
- Logojan A, Molnár-Láng M (2000) Production of *Triticum aestivum*—*Aegilops biuncialis* chromosome additions. *Cereal Res Commun* 28:221–228
- Lukaszewski AJ (1988) A comparison of several approaches in the development of disomic addition lines of wheat. In: Miller TE, Koebner RMD (eds) *Proc 7th Int Wheat Genet Symp* Cambridge, UK, pp 363–367
- Lukaszewski AJ (1991) Development of aneuploid series in hexaploid triticale. In: Baier A (ed) *Proc 2nd Int Tritical Symp PASSO Fundo, Brazil* pp 397–400
- Lukaszewski AJ (2000) Manipulation of the 1RS.1BL translocation in wheat by induced homoeologous recombination. *Crop Sci* 40:216–225. doi:10.2135/cropsci2000.401216x
- Lukaszewski AJ, Gustafson JP (1983) Translocations and modifications of chromosomes in triticale × wheat hybrids. *Theor Appl Genet* 64:239–248. doi:10.1007/BF00303771

- Maan SS (1976) Cytoplasmic homology between *Aegilops squarrosa* L. and *Ae. cylindrica* Host. *Crop Sci* 16:757–761. doi:10.2135/cropsci1976.0011183X001600060004x
- Makkouk KM, Comeau A, Ghulam W (1994) Resistance to barley yellow dwarf luteovirus in *Aegilops* species. *Can J Plant Sci* 74:631–634. doi:10.4141/cjps94-113
- Marais GF, Horn M, Du Toit F (1994) Intergeneric transfer (rye to wheat) of a gene(s) for Russian wheat aphid resistance. *Plant Breeding* 113:265–271. doi:10.1111/j.1439-0523.1994.tb00735.x
- Mettin D, Bluthner WD, Schlegel G (1973) Additional evidence on spontaneous 1B/1R wheat–rye substitutions and translocation. In: Sears ER, Sears LMS (eds) *Proc 4th Int Wheat Gen Symp*, Mo Agric Exp Stn Columbia, pp 179–184
- Miller TE, Riley R (1972) Meiotic chromosome pairing in wheat-rye combinations. *Genet. Ibér* 24:1–10
- Miller TE, Reader SM, Purdie KA, King IP (1994) Determination of the frequency of wheat-rye chromosome pairing in wheat × rye hybrids with and without chromosome 5B. *Theor Appl Genet* 98:255–258. doi:10.1007/BF00225150
- Molnár I, Benavente E, Molnár-Láng M (2009) Detection of intergenomic chromosome rearrangements in irradiated *Triticum aestivum*/*Aegilops biuncialis* amphiploids by multicolour genomic in situ hybridization. *Genome* 52:156–165. doi:10.1139/G08-114
- Molnár I, Dulai S, Molnár-Láng M (2008) Can the drought tolerance traits of *Ae. biuncialis* manifest even in the wheat genetic background? *Acta Biol Szeged* 52:175–178
- Molnár I, Gáspár L, Sárvári É et al (2004) Physiological and morphological responses to water stress in *Aegilops biuncialis* and *Triticum aestivum* genotypes with differing tolerance to drought. *Funct Plant Biol* 31:1149–1159. doi:10.1071/FP03143
- Molnár I, Cifuentes M, Schneider A, Benavente E, Molnár-Láng M (2011a) Association between SSR-rich chromosome regions and intergenomic translocation breakpoints in natural populations of allopolyploid wild wheats. *Ann Bot* 107:65–76. doi:10.1093/aob/mcq215
- Molnár I, Kubaláková M, Šimková H, Cseh A, Molnár-Láng M, Doležel J (2011b) Chromosome isolation by flow sorting in *Aegilops umbellulata* and *Ae. comosa* and their allotetraploid hybrids *Ae. biuncialis* and *Ae. geniculata*. *PLoS ONE* 6:e27708. doi:10.1371/journal.pone.0027708
- Molnár I, Šimková H, Leverington-Waite M, Goram R, Cseh A, Vrána J, Farkas A, Doležel J, Molnár-Láng M, Griffiths S (2013) Syntenic relationships between the U and M genomes of *Aegilops*, wheat and the model species *Brachypodium* and rice as revealed by COS markers. *PLoS ONE* 8:e70844. doi:10.1371/journal.pone.0070844
- Molnár I, Schneider A, Molnár-Láng M (2005) Demonstration of *Aegilops biuncialis* chromosomes in a wheat background by genomic in situ hybridization (GISH) and identification of U chromosomes by FISH using GAA sequences. *Cereal Res Commun* 33:673–680. doi:10.1556/CRC.33.2005.2-3.134
- Molnár-Láng M, Sutka J (1994) The effect of temperature on seed set and embryo development in reciprocal crosses of wheat and barley. *Euphytica* 78:53–58. doi:10.1007/BF00021397
- Molnár-Láng M, Cseh A, Szakács É, Molnár I (2010) Development of a wheat genotype combining the recessive crossability alleles *kr1kr1kr2kr2* and the 1BL.1RS translocation, for the rapid enrichment of 1RS with new allelic variation. *Theor Appl Genet* 120:1535–1545. doi:10.1007/s00122-010-1274-0
- Molnár-Láng M, Kruppa K, Cseh A et al (2012) Identification and phenotypic description of new wheat - six-rowed winter barley disomic additions. *Genome* 55:302–311. doi:10.1139/G2012-013
- Molnár-Láng M, Linc G, Nagy ED et al (2002) Molecular cytogenetic analysis of wheat-alien hybrids and derivatives. *Acta Agron Hung* 50:303–311. doi:10.1556/AAgr.50.2002.3.8
- Molnár-Láng M, Linc G, Sutka J (1996) Transfer of the recessive crossability allele *kr1* from Chinese Spring into the winter wheat variety Martonvásári 9. *Euphytica* 90:301–305. doi:10.1007/BF00027480
- Molnár-Láng M, Linc G, Logojan A, Sutka J (2000a) Production and meiotic pairing behaviour of new hybrids of winter wheat (*Triticum aestivum*) × winter barley (*Hordeum vulgare*). *Genome* 43:1045–1054. doi:10.1139/g00-079

- Molnár-Láng M, Linc G, Friebe BR, Sutka J (2000b) Detection of wheat-barley translocations by genomic in situ hybridization in derivatives of hybrids multiplied in vitro. *Euphytica* 112:117–123. doi:10.1023/A:1003840200744
- Molnár-Láng M, Novotny C, Linc G, Nagy DE (2005) Changes in the meiotic pairing behaviour of a winter wheat-winter barley hybrid maintained for a long term in tissue culture, and tracing the barley chromatin in the progenies using GISH and SSR markers. *Plant Breeding* 124:247–252. doi:10.1111/j.1439-0523.2005.01097.x
- Molski BA, Luckzak W, Zych J (1985) Protein quantity and quality in rye collections and in agricultural production in Poland. In: EUCARPIA meeting of the cereal section on rye. Svalof, Sweden, pp 491–523
- Mukade K, Kamio M, Hosoda K (1970) The transfer of leaf rust resistance from rye to wheat by intergeneric addition and translocation. *Gamma Field Symp. No. 9. Mutagenesis in Relation to Ploidy Level.* pp 69–87
- Mukai Y, Friebe B, Gill BS (1992) Comparison of C-banding patterns and in situ hybridization sites using highly repetitive and total genomic rye DNA probes of ‘Imperial’ rye chromosomes added to ‘Chinese Spring’ wheat. *Jpn J Genet* 67:71–83. doi:10.1266/jjg.67.71
- Mukai Y, Nakahara Y, Yamamoto M (1993) Simultaneous discrimination of the three genomes in hexaploid wheat by multicolor fluorescence in situ hybridization using total genomic and highly repetitive DNA probes. *Genome* 36:489–494. doi:10.1139/g93-067
- Mullan DJ, Mirzaghaderi G, Walker E et al (2009) Development of wheat-*Lophopyrum elongatum* recombinant lines for enhanced sodium ‘exclusion’ during salinity stress. *Theor Appl Genet* 119:1313–1323. doi: 10.1007/s00122-009-1136-9
- Nagy DE, Linc G, Molnár-Láng M (1998) Molecular cytogenetic analysis of a new wheat-rye hybrid with C-banding and genomic in situ hybridization (GISH). (Új búza-rozs amfidiploid molekuláris genetikai elemzése C-sávózással és genomikus in situ hibridizációval (GISH). Abstract in English). *Növénytermelés* 3:253–260
- Nagy DE, Molnár-Láng M, Linc G, Láng L (2002) Identification of wheat-barley translocations by sequential GISH and two-colour FISH in combination with the use of genetically mapped barley SSR markers. *Genome* 45:1238–1247. doi:10.1139/g02-068
- Nagy DE, Eder C, Molnár-Láng M, Lelley T (2003) Genetic mapping of sequence specific PCR-based markers in the short arm of the 1BL.1RS wheat-rye translocation. *Euphytica* 132:243–250. doi:10.1023/A:1025002919746
- Nakai Y (1981) D genome donors for *Aegilops cylindrica* (CCDD) and *Triticum aestivum* (AABBDD) deduced from esterase isozyme analysis. *Theor Appl Genet* 60:11–16. doi:10.1007/BF00275172
- Nasuda S, Friebe B, Busch W et al (1998) Structural rearrangement in chromosome 2M of *Aegilops comosa* has prevented the utilization of the Compair and related wheat-*Ae. comosa* translocations in wheat improvement. *Theor Appl Genet* 96:780–785. doi:10.1007/s001220050802
- Oliver RE, Xu SS, Stack RW et al (2006) Molecular cytogenetic characterization of four partial wheat-*Thinopyrum ponticum* amphiploids and their reaction to *Fusarium* head blight, tan spot, and *Stagonospora nodorum* blotch. *Theor Appl Genet* 112:1473–1479. doi:10.1007/s00122-006-0250-1
- O’Mara JG (1940) Cytogenetic studies on *Triticeae*. I. A method for determining the effect of individual *Secale* chromosomes on *Triticum*. *Genetics* 25:401–408
- O’Mara JG (1947) The substitution of a specific *Secale cereale* chromosome for a specific *Triticum aestivum* chromosome. *Genetics* 32:99–100
- Pardue ML, Gall JG (1970) Chromosomal localization of mouse satellite DNA. *Science* 168:1356–1358. doi:10.1126/science.168.3937.1356
- Pedersen C, Linde-Laursen I (1994) Chromosomal location of four minor rDNA loci and a marker microsatellite sequence in barley. *Chromosome Res* 2:65–71. doi:10.1007/BF01539456
- Rabinovich SV (1998) Importance of wheat-rye translocations for breeding modern cultivars of *Triticum aestivum* L. *Euphytica* 100:323–340. doi:10.1023/A:1018361819215

- Rajaram S, Villareal R, Mujeeb-Kazi A (1990) Global impact of 1B/1R spring wheats. In: Agronomy Abstracts, ASA, Madison, WI, USA, p. 105
- Rao MVP (1978) The transfer of alien genes for stem rust resistance to durum wheat. In: Ramanujam (ed) Proc 5th Int Wheat Genet Symp New Delhi, India. pp 338–341
- Raupp WJ, Amri A, Hatchett JH et al (1993) Chromosomal location of Hessian fly-resistance genes *H22*, *H23* and *H24* derived from *Triticum tauschii* in the D genome of wheat. *J Heredity* 84:142–145
- Raupp WJ, Gill BS, Friebe B, Wilson DL, Cox TS, Sears RG (1995) The Wheat Genetics Resource Center: germ plasm conservation, evaluation and utilization. In: Li ZS, Xin ZY (eds) Proc 8th Int Wheat Genet Symp, China Agricultural Sciencetech Press, Beijing, China pp 469–475
- Ren TH, Chen F, Yan BJ et al (2011) Genetic diversity of wheat–rye 1BL.1RS translocation lines derived from different wheat and rye sources. *Euphytica* (online) doi 10.1007/s10681-011-0412-3
- Ribeiro-Carvalho C, Guedes-Pinto H, Harrison G, Heslop-Harrison JS (1997) Wheat-rye chromosome translocations involving small terminal and intercalary rye chromosome segments in the Portuguese wheat landrace Barbela. *Heredity* 78:539–546. doi:10.1038/hdy.1997.84
- Riley R, Chapman V (1967) The inheritance in wheat of crossability with rye. *Genet Res Cambridge* 9:259–267. doi:10.1017/S0016672300010569
- Sánchez-Morán E, Benavente E, Orellana J (1999) Simultaneous identification of A, B, D and R genomes by genomic in situ hybridization in wheat–rye derivatives. *Heredity* 83:249–252. doi:10.1038/sj.hdy.6885570
- Sarma NP, Natarajan AT (1973) Identification of heterochromatic regions in the chromosomes of rye. *Hereditas* 74:233–237. doi:10.1111/j.1601-5223.1973.tb01124.x
- Sax K, Sax MJ (1924) Chromosome behaviour in a genus cross. *Genetics* 9:454–464
- Schlegel R (1997) Current list of wheats with rye introgressions of homoecologous group I. *Wheat Inf Serv* 84:64–69
- Schlegel R (2006) Rye (*Secale cereale* L.): a younger crop plant with a bright future. In: Sing RJ, Jauhar P (eds) Genetic resources, chromosome engineering, and crop improvement: vol. II cereals. CRC Press, Boca Raton, pp 365–394
- Schlegel R, Korzun V (1997) About the origin of 1RS.1BL wheat-rye chromosome translocations from Germany. *Plant Breeding* 116:537–540. doi:10.1111/j.1439-0523.1997.tb02186.x
- Schlegel R, Kynast R, Schwarzacher T et al (1993) Mapping of genes for copper efficiency in rye and the relationship between copper and iron efficiency. *Plant Soil* 154:61–65. doi:10.1007/BF00011072
- Schneider A, Molnár I, Molnár-Láng M (2008) Utilisation of *Aegilops* (goatgrass) species to widen the genetic diversity of cultivated wheat. *Euphytica* 163:1–19. doi:10.1007/s10681-007-9624-y
- Schubert I, Shi F, Fuchs J, Endo TR (1998) An efficient screening for terminal deletions and translocations of barley chromosomes added to common wheat. *Plant J* 14:489–495. doi:10.1046/j.1365-313X.1998.00125.x
- Schwarzacher T, Leitch AR, Bennett MD, Heslop-Harrison JS (1989) In situ localization of parental genomes in a wide hybrid. *Ann Bot* 64:315–324
- Schwarzacher T, Anamthawat-Jónsson K, Harrison GE et al (1992) Genomic in situ hybridization to identify alien chromosomes and chromosome segments in wheat. *Theor Appl Genet* 84:778–786. doi:10.1007/BF00227384
- Sears ER (1956) The transfer of leaf rust resistance from *Aegilops umbellulata* to wheat. *Brookhaven Symp Biol* 9:1–22
- Sears RG, Hatchett JH, Cox TS, Gill BS (1992) Registration of Hamlet, a Hessian fly resistant hard red winter wheat germplasm. *Crop Sci* 32:506. doi:10.2135/cropsci1992.0011183X003200020061x
- Schneider A, Linc G, Molnár I, Molnár-Láng M (2005) Molecular cytogenetic characterization of *Aegilops biuncialis* and its use for the identification of five derived wheat/*Aegilops biuncialis* disomic addition lines. *Genome* 48:1070–1082. doi:10.1139/G05-062

- Sepsi A, Molnár I, Szalay D, Molnár-Láng M (2008) Characterization of a leaf rust resistant wheat-*Th. ponticum* partial amphiploid BE-1, using sequential multicolor GISH and FISH. *Theor Appl Genet* 116:825–834. DOI:10.1007/s00122-008-0716-4
- Sepsi A, Molnár I, Molnár-Láng M (2009) Physical mapping of a 7A.7D translocation using multicolour genomic in situ hybridization and microsatellite marker analysis. *Genome* 52:748–754
- Shaked H, Kashkush K, Ozkan H et al (2001) Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* 13:1749–1759. doi:10.1105/tpc.13.8.1749
- Sharma D, Knott D (1966) The transfer of leaf rust resistance from *Agropyron* to *Triticum*. *Can J Genet Cytol* 8:137–143. doi:10.1139/g66-018
- Shepherd KW (1973) Homeology of wheat and alien chromosomes controlling endosperm protein phenotypes. In: Sears ER, Sears LMS (eds), *Proc 4th Int Wheat Genet Symp*, Columbia, Missouri, USA, pp 745–760
- Shepherd KW, Islam AKMR (1981) Wheat:barley hybrids—the first eighty years. In: Evans LT, Peacock WJ (eds) *Wheat Science—today and tomorrow*. Cambridge University Press, Cambridge, pp 107–128
- Shepherd KW, Islam AKMR (1988) Fourth compendium of wheat-alien chromosome lines. In: Miller TE, Koebner RMD (eds) *Proc 7th Int Wheat Genet Symp*, Cambridge pp 1373–1395
- Singh RJ, Jauhar PP (eds) *Genetic resources, chromosome engineering, and crop improvement*. CRC Press, Taylor and Francis, Boca Raton, Florida, USA, pp 365–394. doi:10.1201/9780203489260.ch12
- van Slageren MW (1994) *Wild wheats: a monograph of Aegilops L. and Amblyopyrum* (Jaub. & Spach) Eig (Poaceae). Agricultural University, Wageningen; International Center for Agricultural Research in Dry Areas, Aleppo, Syria
- Snape JW, Parker BB, Leckie D (1987) Progress report. Intervarietal transfer of crossability genes. In: Sutka J, Worland AJ (eds) *Proc EWAC Conference*, EWAC Newsletter, Agric Res Inst Hung Acad Sci, Martonvásár, Inst Plant Sci Res, Cambridge Laboratory pp 17–18
- Somers DJ, Isaak P, Edwards K (2004) A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* 109:1005–1014. doi:10.1007/s00122-004-1740-7
- Szakács É, Molnár-Láng M (2007) Development and molecular cytogenetic identification of new winter wheat/winter barley (Martonvásári 9 kr1/Igri) disomic addition lines. *Genome* 50:43–50. doi:10.1139/g06-134
- Szakács É, Molnár-Láng M (2010a) Identification of new winter wheat—winter barley addition lines (6HS and 7H) using fluorescence in situ hybridization and the stability of the whole ‘Martonvásári 9 kr1’—‘Igri’ addition set. *Genome* 53:35–44. doi:10.1139/G09-085
- Szakács É, Molnár-Láng M (2010b) Molecular cytogenetic evaluation of chromosome instability in *Triticum aestivum*—*Secale cereale* disomic addition lines. *J Appl Gen* 51:149–152. doi:10.1007/BF03195723
- Szalay D (1979) Faj- és nemzetséghibridek felhasználása a búzanemesítésben (Use of interspecific and intergeneric hybrids in wheat breeding). In: Bálint A (ed) *A búza jelene és jövője* (The present and future of wheat). Mezőgazdasági Kiadó, Budapest, pp 61–66
- Thomas J, Chen Q, Talbert L (1998) Genetic segregation and the detection of spontaneous wheat-alien translocations. *Euphytica* 100:261–267. doi:10.1023/A:1018320710129
- Tixier MH, Sourdille P, Charmet G et al (1998) Detection of QTLs for crossability in wheat using a doubled haploid population. *Theor Appl Genet* 97:1076–1082. doi: 10.1007/s001220050994
- Tsitsin NV (1960) The significance of wide hybridization in the evolution and production of new species and forms of plants and animals. in Tsitsin NV (ed) *Wide hybridization in plants*. Jerusalem: Israel Program for Science (Translation) pp 2–30
- Tsunewaki K (1964) Genetic studies of a 6 × -derivative from an 8 × -Triticale. *Can J Genet Cytol* 6:1–11. doi:10.1139/g64-001
- Villareal RL, Banuelos O, Mujeeb-Kazi A, Rajaram S (1998) Agronomic performance of chromosome 1B and T1BL.1RS, near-isolines in the spring bread wheat Seri M82. *Euphytica* 103:195–202. doi:10.1023/A:1018392002909

- Wang RRC (2011) *Agropyron* and *Psathyrostachys*. In: Kole C (ed) Wild crop relatives: genomic and breeding resources. Cereals. Springer-Verlag, Berlin Heidelberg, pp 1–76. doi:10.1007/978-3-642-14228-4_1
- Wang RRC, Jensen KB (2009) Wheatgrasses and wildryes, Chap. 3. In: Singh RJ (ed) Genetic resources, chromosome engineering and crop improvement, vol 5, Forage Crop. CRC, Boca Raton USA, pp 42–79
- Wang ZN, Hang A, Hansen J et al (2000) Visualization of A- and B-genome chromosomes in wheat (*Triticum aestivum* L.) × jointed goatgrass (*Aegilops cylindrica* Host) backcross progenies. Genome 43:1038–1044. doi:10.1139/g00-080
- Wojciechowska B, Pudelska H (1993) Hybrids from reciprocal barley-wheat crosses. Gen Polonica 34:1–13
- Zeller FJ (1973) 1B/1R chromosome substitutions and translocation. In: ER S, LMS S (eds) Proc 4th Int Wheat Genet Symp, Columbia, Missouri, USA pp 209–221
- Zeller FJ, Fuchs E (1983) Cytologie und Krankheitsresistenz einer 1A/1R- und mehrerer 1B/1R-Weizen-Roggen-Translokationsorten. Z. Pflanzenzücht 90:285–296
- Zeven AC (1987) Crossability percentages of some 1,400 bread wheat varieties and lines with rye. Euphytica 36:299–319. doi:10.1007/BF00730677
- Zhong G, McGuire PE, Qualset CO, Dvorak J (1994) Cytological and molecular characterization of a *Triticum aestivum*/*Lophopyrum ponticum* backcross derivative resistant to barley yellow dwarf. Genome 37:876–881(1994). <http://dx.doi.org/10.1139>

Chapter 12

Radiation Hybrids: A valuable Tool for Genetic, Genomic and Functional Analysis of Plant Genomes

Ajay Kumar, Filippo M. Bassi, Monika K. Michalak de Jimenez,
Farhad Ghavami, Mona Mazaheri, Kristin Simons, Muhammad J. Iqbal,
Mohamed Mergoum, Shahryar F. Kianian and Penny M. A. Kianian

Contents

12.1	Effects of Radiation on Plant Genomes	287
12.1.1	Considerations on the Radiation Effect for Mutant Population Development ..	288
12.2	Application of Radiation Mutagenesis in Crop Improvement	290
12.2.1	Mutation Breeding Contribution to Crop Improvement	290
12.2.2	Advantages and Disadvantages of Mutation Breeding	291
12.3	Radiation Hybrid Mapping of Genomes	293
12.3.1	Why Radiation Hybrid Mapping?	293
12.3.2	Radiation Hybrid Mapping in Animal Systems	294
12.3.3	Radiation Hybrid Mapping in Plants: Limited Number of Studies with Important Implications	295
12.3.4	RH Mapping Panel Development in Plants	299
12.3.5	Potential of Developing RH Panels for Any Plant Species	303
12.3.6	Applications and Prospects for Radiation Hybrids in Plants	304
References	311

S. F. Kianian (✉)

USDA-ARS Cereal Disease Laboratory, University of Minnesota,
St. Paul, MN 55108, USA

e-mail: Shahryar.Kianian@ars.usda.gov

A. Kumar · M. K. M. de Jimenez · M. Mazaheri ·

K. Simons · M. J. Iqbal · M. Mergoum

Department of Plant Sciences, North Dakota State University,
Fargo, ND 58108, USA

F. Ghavami

Department of Plant Pathology, University of Minnesota,
St. Paul, MN 55108, USA

P. M. A. Kianian

Department of Horticultural Science, University of Minnesota,
St. Paul, MN 55108, USA

F. M. Bassi

Durum Wheat Breeding, ICARDA, Rabat, Morocco

Abstract Radiation has been used as a mean to break and transfer fragments of DNA from one plant species to another. Early examples include the experiments by Sears, (Brookhaven Symp Biol 9:1–22, 1956) to transfer rust resistance genes from *Aegilops umbellulata* to wheat. Radiation found its niche as a mutagen due to advances in nuclear technology and formation of the International Atomic Energy Agency and their sponsorship of developing mutation breeding through “Mutation Enhanced Technologies for Agriculture”. Mutation breeding has resulted in the release of several important cultivars. Although radiation was used in plants for the mutation and introgression of genes from related species (Sears, Brookhaven Symp Biol 9:1–22, 1956; Driscoll and Jensen, Genetics 48:459–468, 1963; Riley and Law, Stadler Genet Symp 16:301–322, 1984; Sears, Crop Sci 33:897–901, 1993), this approach was not used for mapping. This aspect of radiation application was first utilized in animal cell culture lines to generate radiation hybrid (RH) panels. In the beginning these panels were generated for single chromosomes but evolved to the development of whole genome panels. This technology matured in animal systems with the onset of genomics era by its use in the development of high resolution RH-based physical maps for many species before or during the development of complete genome sequence information. The advantages of this system are: (1) radiation-induced breaks are independent of recombination events providing higher and more uniform resolution, (2) radiation dosage could be adjusted to provide varied resolution without greatly affecting the population size and (3) all markers regardless of their polymorphism can be mapped on RH panels. Plant scientists followed these studies by the development of RH panels for individual chromosomes or whole genomes. However, early RH panels in plant systems were of low to medium resolution and of limited use in physical mapping. Recently, RH panels have been produced resulting in map resolutions of 200–400 Kb. These high resolution panels promise the same value as animal systems in helping generate a complete genome sequence with a fraction of the cost of traditional methods. But the use of radiation in plants has matured to go beyond physical mapping by its application to gene cloning and forward/reverse genetic studies. These applications take advantage of plasticity offered by many plant species in tolerating radiation to produce seed and live progeny. This ability allows scientists to phenotype RH lines and to associate the phenotypic data with the genotypic data. The great potential of this system is just being realized.

Genetic variation is the key to phenotypic improvement in plants. Variation in the genome provides individuals the potential to nucleate and interact with DNA to adapt to changing environmental pressures and improve their chances of survival. Genetic variation also lays the foundation of genomic studies and efforts made towards crop improvement. Genetic variability in any population can be the result of natural processes such as recombination, mistakes in DNA replication and repair, gene flow or induced mutagenesis which results from chemical or radiation treatment. In this chapter, we discuss the use of radiation-induced changes on genes and chromosomes to understand the structure and function of plant genomes, dissect genetic mechanisms, and the potential use of these changes for plant improvement.

Keywords Radiation hybrids · Mutation · Deletion · High-resolution physical map · Forward and reverse genetics · Gene cloning

12.1 Effects of Radiation on Plant Genomes

A clear understanding of the mechanisms governing radiation-mediated damage of DNA and its subsequent repair is important for successful application in analyzing genome structure and function. The amount of damage caused by ionizing radiation as a result of decay of radioactive atoms is a direct function of their energy level and ability to penetrate biological material (Argonne National Laboratory 2005). During exposure of biological tissue to radioactive isotopes, the electromagnetic waves of bundles (quanta) of energy travel across the tissue's cells with a complete lack of target specificity. The probability of any cell component to interact with ionizing energy is proportional to the fraction of tissue that they occupy. Since water accounts for over 80 % of the volume of any metabolically active plant cell, water radiolysis is the primary effect of radiation in plant tissues (Britt 1996). Ionization of water gives rise to reactive oxygen species (ROS) such as hydrogen peroxide (H_2O_2), superoxide anion (O_2^-), hydroxyl radicals (OH), and singlet oxygen (O) (Wi et al. 2007). When ROS originate in the nucleus and interact with DNA, the sugar phosphate backbone undergoes oxidative damage, which inevitably leads to single-strand breaks (SSB; Britt 1996). While SSB are generally repaired in an efficient and error-free fashion, continuous exposure to ROS will eventually cause formation of SSB on both strands, resulting in double-strand breaks (DSB). More rarely, the sugar backbone can be directly hit by an ionizing atom and the absorbed energy might result in SSB. In this case, the sensitization of the opposite un-nicked strand may result in an increased yield of DSB (Britt 1996). Because nucleotides near the ends of breaks are rapidly degraded, DSB generally expand into gaps that cannot simply be rejoined to restore the original sequence (Britt 1999).

Eukaryotic cells immediately respond to the damage by activating the DNA repair mechanisms. Homologous recombination (HR) is an error-free mechanism of DSB repair that involves the use of a complementary template sequence to fill the gap separating the two breaks. For this repair mechanism to be successful, a compatible template sequence is required in close proximity of the breaks to be repaired. This is the case when sister chromatids are brought together to form the synaptonemal complex during leptotene to early pachytene phases. It has been suggested that eukaryotic cells employ the HR mechanism to repair DSB only during this specific portion of the cell-cycle (Ahmed et al. 2010). During any other stages of the cell life DSB are repaired by non-homologous end joining (NHEJ). This mechanism, often referred to as "illegitimate recombination", employs a complex of proteins to bring together DNA termini and joins those using micro-homologies of two to five bases to stabilize the ligation reaction (Britt 1999; Weterings and Gent 2004; Sengupta and Harris 2005). Because there is no mechanism to ensure pairing of the two original chromosome ends, NHEJ is an error-prone mechanism which can lead to deletions,

chromosomal inversions, translocations and partial duplications (Britt 1999). More energetic or longer radiation exposures generally lead to a higher number of breaks, which in turn increase the frequency of DNA repair and larger aberrations (Hlatky et al. 2002).

12.1.1 Considerations on the Radiation Effect for Mutant Population Development

The mechanism of radiation induced break and repair is exploited in studies to generate a binary polymorphism (0-deletion vs. 1-retention) at the DNA level. Physical or molecular markers can be employed to identify this binary polymorphism and DNA to DNA, or DNA to phenotype associations established based on relative co-retention frequencies. The lack of target specificity for radiation damage results in sequence/chromatid independent DNA damage. Thus, DSBs on each chromatid are formed and repaired independently. As a result, two neighboring cells may have different mutations, and within each cell sister chromatids may have different aberrations resulting in chimeric tissue and plants. Thus, an M_1 individual will have mutagenized loci in a hemizygous state. At meiosis, these loci segregate and the ratio will depend on the genetically effective cell number (GECN) of the species. Information regarding GECN is important in any radiation-induced mutation experiment for it allows calculation of the rate of recovery of a given mutant. The model plant *Arabidopsis thaliana* has a $GECN = 2$ (Page and Grossniklaus 2002). Hence, after selfing of an irradiated *Arabidopsis* M_1 line, a segregation ratio of 5 normal: 2 hemizygous: 1 homozygous mutant is expected. GECN has been calculated for a few plant species and the information available to date is summarized in Table 12.1. The short list of plant species for which GECN is currently known indicates how this aspect is often underestimated when mutant populations are developed.

In order to maximize the genetic variability of a mutant population, the number of aberrations generated for each individual should be maximized. Based on the principles of radiation damage discussed earlier, four strategies can be considered: (1) irradiate soft tissue with limited resistance to energy penetration, (2) maximize the amount of water present within the tissue, (3) increase radiation dosages (longer exposure time and shorter distance from the radioactive source), and/or (4) use high energy radiation sources. In Fig. 12.1, *Triticum aestivum* M_3 progenies from an M_2 line, derived from γ -ray treatment of the boron toxicity tolerant cultivar 'Halberd' are shown. The hemizygous M_2 line was identified to be mutant for the boron toxicity tolerance locus (*Bo1*; Schnurbusch et al. 2007), as shown by average length of roots after treatment in boron toxic media. After selfing of this mutant line, its M_3 progenies segregate for the *Bo1* locus in a 1:2:1 ratio with long (homozygous), average (hemizygous), and short (nullizygous) root length after exposure to toxic boron (courtesy of Dr. Peter Langridge, Australian Center for Plant Functional Genomics, Adelaide, Australia).

Table 12.1 Genetically effective cell number (GECN) in some plant species

Species	GECN	Reference
<i>Arabidopsis thaliana</i>	2	Page and Grossniklaus 2002
<i>Glycine Max</i>	2	Carroll et al. 1988
<i>Nicotiana plumbaginifolia</i>	2	Blonstein et al. 1991a, b; Pelsy et al. 1991
<i>Medicago sativa</i>	3	Le Signor et al. 2009
<i>Linum usitatissimum</i>	4	Bertagne-Sagnard et al. 1996
<i>Zea mays</i>	4 or more	Anderson et al. 1949; Johri and Coe 1983
<i>Triticum aestivum</i>	5 or more	Kumar et al. 2012b
<i>Hordeum vulgare</i>	6	Hodgdon et al. 1981

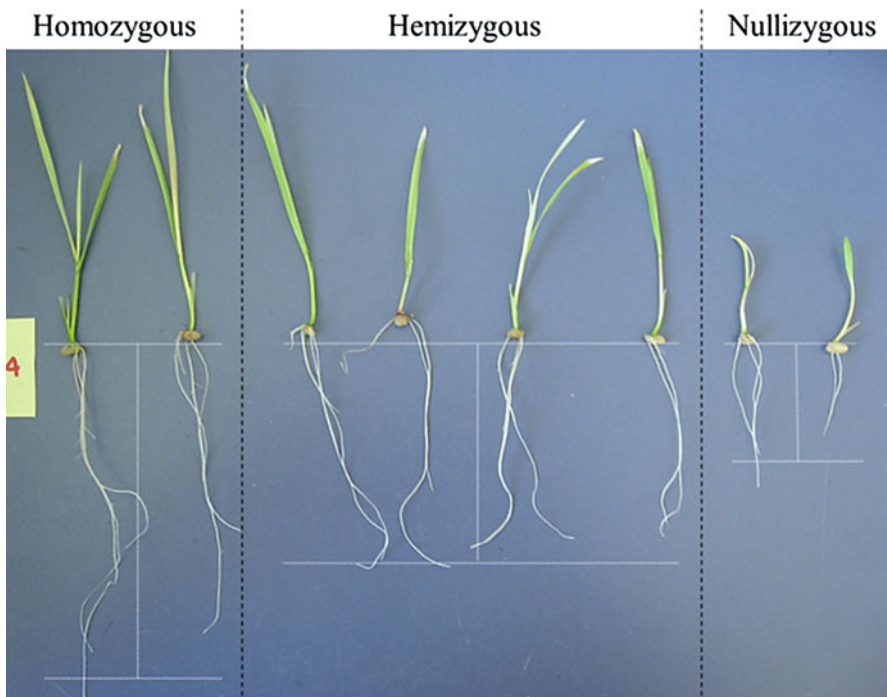


Fig. 12.1 Mendelian segregation of M_3 progenies derived from the selfing of a M_2 *T. aestivum* var. Halberd line hemizygous for the *Bo1* toxicity tolerance locus (Bo1; Schnurbusch et al. 2007), illustrated by the segregation in roots length after exposure to toxic levels of boron. The M_3 progenies segregate for the *Bo1* locus in a 1:2:1 ratio with long (homozygous), average (hemizygous) and short (nullizygous) roots length after exposure to toxic boron media. (Fig. 12.1 kindly provided by Peter Langridge at ACPFG)

12.2 Application of Radiation Mutagenesis in Crop Improvement

Application of X- and γ -rays to induce heritable mutations was suggested as early as 1904 by Hugo de Vries with the potential to generate agronomically desirable phenotypes (Blakeslee 1935). Twenty-four years later, Hermann J. Muller, the founder of mutation genetics and a Nobel Prize winner, in experiments on *Drosophila melanogaster*, demonstrated that mutations can be artificially induced and “may prove of increasing practical use in plant and animal improvement, in the service of man” (Lönnig 2005). Soon after, plant breeders in Sweden, Germany, and Russia utilized radiation-induced mutations in cultivated plants. The first mutants with improved agronomic potential were generated in barley in 1934–1935, named *erectoides* for their characteristic dense ears and stiff straws (Gustafsson 1954). Simultaneously, in 1934 in Indonesia, X-ray-induced tobacco variety ‘Clorina F1’ was released, followed by the cotton variety ‘M.A.9’ in 1948 in India (Maluszynski et al. 2000). In Germany, a mildew-resistant line and 92 additional mutants were reported in barley (Freisleben and Lein 1944). Shebeski and Lawrence (1954) reported barley mutations for stem rust resistance and stiffer straw. Around the same time, stem rust resistance was induced in an oat variety (Konzak 1954). Useful mutations were also induced in other crops including wheat (Oltmann 1950), flax (Hoffmann and Zoschke 1955), and soybean (Zacharias 1956).

12.2.1 Mutation Breeding Contribution to Crop Improvement

According to FAO/IAEA Mutant Variety Database (MVD), a mutant variety is defined as a variety generated by (1) the use of a mutant line developed by physical or chemical mutagenesis, or somaclonal variation; (2) an indirect use of a mutant line as a parental variety in a cross; (3) the use of mutant gene allele; or (4) irradiation-induced translocation of genes from wild species into plant genomes. As of 2011, there are 3,212 mutant plant varieties released worldwide with a vast majority (89 %) being generated through radiation rather than chemical mutagens (Maluszynski et al. 2000; <http://mvgs.iaea.org/AboutMutantVarieties.aspx>). Among the radiation types employed for mutagenesis, γ -rays were the favorite (64 %), followed by X-rays (22 %). Among the mutant varieties, 75 % were crops and 25 % were ornamental and decorative plants. A total of 1,603 crop mutant varieties were released in seed-propagated species, 1,072 in cereals, and 311 in legumes. Mutant varieties have been released so far in more than 50 countries with China, India, the former USSR, Netherlands, Japan, and USA as leaders in the application of induced-mutation breeding (Maluszynski et al. 1995).

A recent (2005) wheat mutant variety ‘Guinness/1322’ had higher yield, drought tolerance, and improved resistance to lodging and shattering compared to its M_0 parent (<http://mvgs.iaea.org/AboutMutantVarieties.aspx>). This variety was derived by γ -radiation of seeds of bread wheat (*T. aestivum* L.) cultivar ‘Katya’ in Bulgaria.

Table 12.2 Economic impact of mutant varieties (modified from Ahloowalia et al. 2004)

Cereals	Country	Mutant variety	Basis of value assessment	Value or area
Rice	Thailand	RD6 and RD15	Total crop value at farm gate for the period 1989–1998	US\$ 16.9 billion
	China	Zhefu 802	Cumulative planted area between 1986–1994	10.6 million ha
	Japan	18 varieties	Total crop value in 1997	US\$ 937 million
	India	PNR-102 and PNR-381	Annual crop value	US\$ 1,748 million
	Australia	Amaroo	Current annual planted area	60–70 % of rice growing area in Australia
	Costa Rica	Camago 8	Current annual planted area	30 % of rice growing area in Costa Rica
	Vietnam	TNDB100 and THDB	Total planted area in 1999	220,000 ha
	Myanmar	Shwewartun	Total planted area in 1993	800,000 ha
Bread wheat	Pakistan	Jauhar 78, Soghat 90 and Kiran 95	Additional income to farmers during 1991–1999	US\$ 87.1 million
Durum wheat	Italy	Creso	Additional income to farmers during 1983–1993	US\$ 1.8 billion
Barley	UK-Scotland	Golden promise	Crop value 1977–2001	US\$ 417 million
	Numerous European countries	Diamant and derived varieties	Area planted in 1972	2.86 million ha

The economic impact of mutation breeding is difficult to assess. A rough estimate of the trade value of the mutant varieties currently cultivated for some of the major cereals is summarized in Table 12.2 modified from Ahloowalia et al. (2004).

12.2.2 *Advantages and Disadvantages of Mutation Breeding*

The most important advantage of induced-mutation breeding is that it results in new variations that can be integrated into existing breeding programs for cultivar improvement. Another advantage is that any cultivar, including modern cultivars, can be used for treatment leading to the creation of new lines that could carry desirable alleles. Reduced vitality and lethality are common shortfalls of plant mutagenesis. However, mutations with negative effects such as delayed maturity, increased

Table 12.3 Examples of repetitive appearance of certain types of barley mutants compiled according to data published by Lundqvist (1986)

Mutant	Appeared	Number of gene loci
Erectoides (dense spike mutants)	205 times	26
Praematurum (early maturity mutants)	110 times	9
Eceriferum (waxless mutants)	1527 times	76
Breviaristatum (short awn mutants)	140 times	17
Exrubrum (anthocyanin-free)	61 times	18
Macrolepis (lemma-like glume mutants)	40 times	1
Hexastichon (six-row)	41 times	1
Intermedium (between two row and six-row)	144 times	11 ^a
Powdery mildew resistant (including all kinds ^b)	154 times	^c

^a 103 of these cases investigated on 11 *int* gene loci

^b 77 mutants were resistant against race D1, 48 had complete resistance, and 29 displayed brown necrosis

^c Of 72 investigated resistant mutants, 54 were found to be distributed on eight genes (the 28 recessive mutants belong to one single locus); for the remaining 18 mutants the number of loci does not appear to be fully established

lodging, or lower grain yield can be overcome by further crossing (Harten 1998). Fortunately, the wide success of mutation breeding and its long history have made this an ethically acceptable procedure in plants, allowing for the production of *in vivo* mutant populations without bureaucratic and moral limitations typically associated with producing transgenic crops or mutants in other species.

One of the disadvantages of mutation breeding is that the majority of mutations generated are recessive and masked by their dominant counterpart. Additionally, compared to classical breeding, mutation breeding is not as effective in improving quantitative traits, since multiple alleles would need to be favorably altered by mutation (Shu 2009). Moreover, there is a physiological maximum to the number of mutations that an organism can tolerate and a finite number of genes that can be mutagenized without causing immediate lethality. This is summarized by the law of recurrent variation, which states that the larger a mutant collection, the less likely is to identify new mutations that are not already present in the collection; in other words, “mutants preferentially arise that already exist” (Table 12.3; Gottschalk 1989; Lönnig 2005). This phenomenon was observed by Swedish breeder Lundqvist who reported around 9,000 barley mutants in 1986.

Another drawback of mutation breeding is a low mutation frequency which makes it crucial to grow and phenotype a large population for selection of desired mutants. However, technological advances as well as plant molecular genetics and genomics tools have helped make mutation breeding more efficient (Shu 2009). In wheat for example, genetic and physical maps, high throughput marker platforms (e.g. Agilent or Illumina), and available sequence data, as well as modern automated plant phenotyping devices (e.g. LemnaTec Scanalyzer 3D unit; <http://www.lemnatec.com/product/scanalyzer-3d-plant-phenotyping>) are examples of changes since the 1950s that could be employed to make mutation breeding a more efficient approach.

12.3 Radiation Hybrid Mapping of Genomes

The first use of radiation in mapping genes was described by Goss and Harris (1975), who used X-ray-induced chromosome breakages to map genes on the human X chromosome. The approach was called RH mapping and in its standard form (in animals), tissue culture cells from a donor species are treated with a lethal dose of radiations (e.g. γ -rays) to induce random chromosomal breaks. These chromosomal fragments are then rescued by fusion with a suitable recipient cell line from a different species. Resulting individual cell lines which contain a fraction of the donor genome retained as a collection of genomic fragments integrated into the recipient genome are referred to as RH lines. A set of approximately 100 RH lines representing the entire donor genome is required to make a mapping panel. Each member of a panel is then analyzed for the presence or absence of DNA markers specific to the donor species. As the radiation-induced breaks are assumed random, the farther apart marker loci are, the higher their chance is to be broken apart and carried on separate chromosomal fragments. By calculating the frequency of co-retention between markers, the distance between them and their order relative to one another is determined, in a manner analogous to meiotic mapping. The RH map unit of distance is centi Rays (cR), which corresponds to one break between two markers in every 100 lines (i.e. similar to centi Morgan).

12.3.1 Why Radiation Hybrid Mapping?

12.3.1.1 Uniform Mapping Resolution Across the Chromosome

In genetic mapping, frequency of recombination—as a result of crossover events—determines the distance between two adjoining loci. However, distribution of meiotic crossover events along the chromosomes is not uniform (Akhunov et al. 2003; Erayman et al. 2004; Mézard 2006; Saintenac et al. 2009). For example, a recent study on wheat chromosome 3B indicated that crossover frequency per physical distance (cM/Mb) within the centromeric segments, as defined by wheat deletion bins, was between 0.006 and 0.012 compared with 0.85 in the sub-telomeric segments (Saintenac et al. 2009). Thus, centromeric regions are characterized by poor recombination frequencies, which pose a clear disadvantage in recombination-based mapping. Deletion bin mapping of a large number of ESTs in wheat indicated that a large number of transcribed genes are located in the centromeric segments (Conley et al. 2004; Hossain et al. 2004a; Linkiewicz et al. 2004; Miftahudin et al. 2004; Munkvold et al. 2004; Peng et al. 2004; Qi et al. 2004; Randhawa et al. 2004). According to some estimates, ~30% of the proximal regions of wheat chromosomes are represented by less than 1% recombination (Erayman et al. 2004). This indicates that recombination mapping will not provide the adequate resolution needed for map-based cloning and physical mapping of many genes in wheat and other plant species. However, RH mapping uses retention/loss frequency as a result of breaks

caused by radiation to generate physical maps that are relatively independent of recombination. These breaks caused by radiation are expected to be random and map resolutions more uniform across the length of a given chromosome.

12.3.1.2 Higher Resolution Without Increasing the Population Size

The resolution in RH mapping is determined by the number of breaks along the chromosome, which can be adjusted by modifying the radiation dosage while developing a panel. Thus, RH panels with higher levels of resolution than those obtained by meiotic mapping may be produced (Stewart et al. 1997) without increasing the population size (< 1 Mb; Weikard et al. 2006; Karere et al. 2008). This property of RH mapping is critical in constructing an accurate, sequence ready, marker scaffold for detailed analysis and sequence assembly of the complex plant genomes. These RH panels can be used to order contigs of overlapping cloned fragments spanning the entire genome or the particular region of interest. Construction of such maps using recombination-based mapping would be difficult due to low resolution and uneven distribution of recombination across the genome.

12.3.1.3 Polymorphism is Not a Requirement for RH Mapping

One of the advantages of RH mapping is that it uses assays for presence and absence of marker loci and does not rely on allelic polymorphism. Therefore, RH panels constitute one of the most expedient methods for high-resolution assignment of ESTs and genes to linkage groups, and provide an efficient tool for inter-species comparative genome analysis (Kynast et al. 2002; Hossain et al. 2004b; Wardrop et al. 2004; Kalavacharla et al. 2006; Weikard et al. 2006).

12.3.2 Radiation Hybrid Mapping in Animal Systems

Since the publication of initial RH maps by Goss and Harris (1975), this method has been widely adopted in various animal systems. In humans, a 100 kb-resolution contiguous map with ~ 41,000 ordered STSs covering 30,000 unique human genes was constructed using the RH mapping approach (Stewart et al. 1997; Deloukas et al. 1998). The success of this approach in human genome studies resulted in its wide application in RH mapping of other animals (Faraut et al. 2009). During the last two decades, RH mapping has been successfully used in a number of mammalian and non-mammalian species (for review, see Faraut et al. 2009). The success of this approach can be measured by the fact that RH maps have played a major role in the whole genome sequencing and assembly of human (International Human Genome Sequencing Consortium 2001) as well as many domestic animals (Lander et al. 2001; Waterston et al. 2002; Gibbs et al. 2004; Lindblad-Toh et al. 2005; Elsik et al. 2009; Wade et al. 2009).

Table 12.4 Marker loss frequency in D-genome RH panel of AL8/78 for five SSR markers belonging to chromosome 4D. (Kumar et al. 2012b)

Chromosome	Marker name	Deletion bin location	Loss frequency (%)
4DS	cfid106	4DS1-0.53-0.67	2.45
4DS	gwm165	C-4DS1-0.53	2.91
4DL	cfid71	C-4DL9-0.31	3.17
4DL	cfid84	C-4DL9-0.31	3.20
4DL	cfid89	4DL12-0.71-1.00	3.00

12.3.3 Radiation Hybrid Mapping in Plants: Limited Number of Studies with Important Implications

Compared with the hundreds of published RH studies in animals, only few have been reported in plants. However, these few studies in plants have yielded interesting results. Before discussing further the key findings, it is important to make a distinction between RH approaches: in animals, RH panels have so far only been produced *in vitro*, where a radiated genome is rescued by fusion with a recipient cell line, hence generating a ‘hybrid’ radiated cell. In plants, this *in vitro* approach has been followed to a limited extent, but the possibility of creating *in vivo* panels has received more attention. In particular, *in vivo* panels are the result of artificial crosses between a radiated species and a second ‘rescuer’ species. As such, the resulting RH cells are not the result of cell fusion but species fusion. The *in vivo* RH lines are F₁ progenies of a cross and as such can be defined as ‘hybrid’, as defined in the first plant RH paper (Riera-Lizarazu et al. 2000).

12.3.3.1 Homogeneous Mapping Resolution Across Genome

One of the main expectations from RH mapping is its uniformity in resolution across the length of a chromosome. The first study by Riera-Lizarazu et al. (2000) based on 33 maize specific DNA markers (12 SSRs and 21 RFLPs), showed that radiation-induced breakage along maize chromosome 9 was homogeneous, and there was no indication of preferential breakage or lack of breakage for a particular chromosomal region. In a study in wheat, where a panel of > 1,500 radiation hybrids for the D-genome were characterized with 35 markers selected across all chromosomes, the marker loss frequency within a chromosome was homogenous except for chromosome 7D, where one marker (barc1046) showed significantly higher loss than others (Kumar et al. 2012b). The five markers present on chromosome 4D, with a better representation of the whole chromosome including centromeric, sub-telomeric, and telomeric regions, showed a homogenous marker loss across all the deletion bins (Table 12.4). Similar results were reported elsewhere in wheat for two chromosomes (5D and 6D) (Riera-Lizarazu et al. 2010), and in cotton (Gao et al. 2006), where no significant differences in marker loss were observed. These studies do not take into account detailed variations on a given chromosome as they have utilized few markers looking across large segments of the genome.

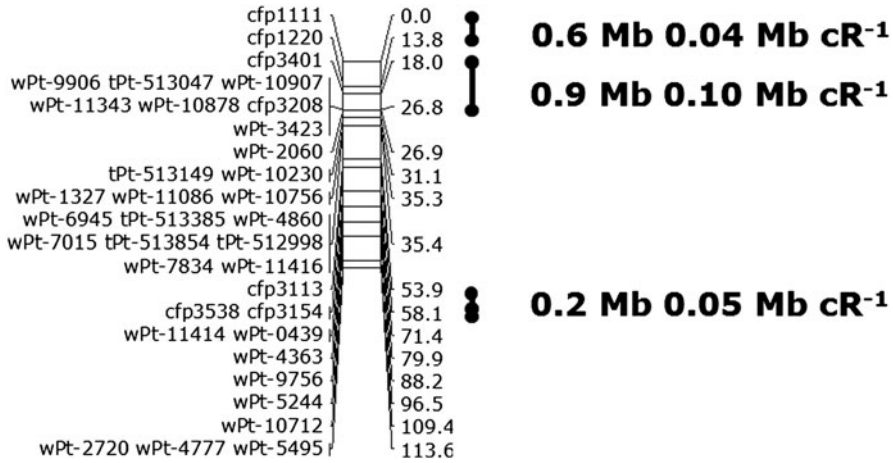


Fig. 12.2 A part of radiation hybrid map of a portion of chromosome 3B of wheat from Kumar et al. (2012a) showing the anchored BAC contigs and cR/Mb resolution within the contigs **a**. In this map markers are presented to the left of the figure with distances in centi-rays (cR) to the right starting with 0.0 to 113.6. Several BAC contigs that were assigned to this segment along with their length in mega-bases (Mb) are indicated in dark lines with filled circles indicating their position. Resolution of this map calculated based on assigned BAC contigs are calculated as Mb cR⁻¹. The average resolution for this portion of chromosome 3B for the RH map was 0.06 Mb cR⁻¹.

A recent study generated a high density RH map of wheat chromosome 3B with 541 marker loci spanning a total distance of 1871.9 cR (Fig. 12.2; Kumar et al. 2012a). Detailed comparisons with a genetic map of similar quality confirmed that (1) the overall resolution of the RH map was 10.5 fold higher and (2) six fold more uniform. A significant interaction ($r = 0.879$ at $P = 0.01$) was observed between the radiation-induced DNA repair mechanism and the distribution of crossing-over events. This observation could be explained by accepting the possibility that the DNA repair mechanism in somatic cells is affected by the chromatin state in a way similar to the effect that chromatin state has on recombination frequencies in gametic cells. Thus, the data presented in this paper supported, for the first time *in vivo* the hypothesis of a non-casual interaction between recombination hot-spots and DNA repair. Despite this significant correlation indicating non-uniform distribution of radiation-induced breakage/repair events across a given chromosome, the RH map of the centromeric region had about 100 fold higher resolution than the genetic map of the same segment (Kumar et al. 2012a). Thus RH maps produce a much higher and more uniform resolution, better representing the physical genome than the genetic maps.

12.3.3.2 Higher Mapping Resolution

Since the RH method does not depend on meiotic recombination and mapping resolution can be improved by increasing the radiation dose, high-resolution RH maps can be generated with fewer individuals as compared with genetic mapping. For

instance, the radiation hybrid map of chromosome 3B mentioned earlier (Kumar et al. 2012a) showed on average 10.5X higher resolution (using 92 RH lines) than a recombination-based genetic map (using ~400 DH lines; Saintenac et al. 2009) with the highest resolution of 136X in the centromeric region of the chromosome (Table 12.5).

The resolution of RH mapping populations depends on chromosomal breaks, which are influenced by radiation dosage. Generating panels with low, medium, and high resolution by altering radiation dose (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2006), will help in development of contiguous maps (Gyapay et al. 1996; Stewart et al. 1997; Olivier et al. 2001). Similar to animal studies (Lunetta et al. 1996; Olivier et al. 2001), in plants an increase in marker loss and chromosome rearrangements has been observed with the increase in γ -radiation dosage (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2006). In cotton, it was shown that 80 Gy RH mapping panel (Gao et al. 2006) offered several significant advantages, including more breakages and more deletion patterns compared to a 50 Gy panel (Gao et al. 2004). However, the tolerance for radiation dosage in plants is limited and chromosomal rearrangements can complicate map analysis. Therefore, optimum dosage where chromosome breaks are maximized while plant survival and seed set are not compromised needs to be determined for the species under investigation.

In whole genome sequencing projects of large and complex genomes, the most successful approach thus far has involved the ordering of BAC contigs into a continuous physical map of the genome. For this purpose mapping population(s) used for ordering needs to provide sufficient resolution to accurately anchor the BAC contigs. This has been achieved for telomeric regions of chromosomes using large mapping populations but segments proximal to the centromeres are hardly accessible by recombination and remain un-resolved. However, data clearly indicates that RH maps provide the needed resolution for the entire genome with much smaller populations to achieve this goal. In a panel of 87 RH lines, Kalavacharla et al. (2006) identified 2,312 chromosome breaks on the chromosome 1D using 378 markers and suggested a ~199 kb/break mapping resolution for that population. In another study, markers designed from the BAC end sequences belonging to a single contig suggested a resolution of < 150 kb in a 1,510 RH line panel for D-genome of wheat (Kumar et al. 2012b). This resolution would be sufficient to anchor most BAC contigs to a physical map. The aforementioned study on chromosome 3B utilizing a population of only 92 lines, allowed the assignment of BAC contigs that were larger than 1.8 Mb in size (Fig. 12.2). The BAC contigs that could not be accurately aligned to the RH map in this study were 0.4 Mb or less in size (Kumar et al. 2012a). These results in plants suggest that the RH mapping approach will greatly aid the physical mapping and sequence assembly efforts and complement other novel techniques currently being developed for use in the production of physical maps of large plant genomes.

12.3.3.3 Marker Retention Frequency in Plant RH Panels

Simulation studies have suggested that a retention/loss frequency of 50% would be optimal for mapping purpose (Jones 1996). However, 50% marker loss has rarely been achieved in plants (for seed and pollen irradiation panels), with the exclusion

Table 12.5 Resolution comparison between RH and recombination-based maps across the 3B cytological stocks

Deletion bins	3BS3- 0.87-1.00	3BS8- 0.78-0.87	3BS9- 0.57-0.75	3BS2- 0.56-0.57	3BS1- 0.33-0.55	C-3BS1- 0.33	C-3BL2- 0.22	3BL2- 0.22-0.28	3BL1- 0.31-0.50	3BL10- 0.50-0.63	3BL7- 0.63-1.00	Chr. ^a
No. of markers	7	155	9	3	44	28	27	29	47	4	187	540
Bin size (Mb)	56	39	78	4	95	142	124	33	39	73	208	992
Marker density (Mb)	8	0.3	8.7	1.3	2.2	5.1	4.6	1.1	0.8	18.3	1.1	4.7
Marker ⁻¹												
RH Map size (cR) ^b	79.7	572.9	69.2	36.7	113.7	115.6	99.1	58.9	37	41.3	459.6	1871.9
Resolution (Mb cR ⁻¹) ^c	0.7	0.1	1.1	0.1	0.8	1.2	1.3	0.6	1.1	1.8	0.5	0.5
Genetic Map size (cM) ^d	5.2	33.1	13.4	-	1.6	0.9	1.6	1.6	6.2	3.5	104.2	179.1
Resolution (Mb cM ⁻¹) ^c	10.8	1.2	5.8	-	61.3	167.1	80	21.3	6.3	21.2	2	5.5
Map size ratio (cR cM ⁻¹)	15.3	17.3	5.2	-	73.4	136	63.9	38	6	12	4.4	10.5

^a Chr., whole chromosome values are not averages for each category but re-calculated for the overall values

^b Kumar et al. (2012a)

^c Resolution, a potential to uniquely resolve two markers at a distance equal or larger than this value

^d Sainetnac et al. (2009)

of the laborious and often impossible option of protoplast cell fusion (Wardrop et al. 2004) or a small fraction of *in vivo* RH panels, where less than 0.5 % of the lines show retention frequencies between 40–60 % (Kumar et al. 2012a, b; Bassi et al. 2013; Michalack de Jimenez et al. 2013). In plants, irradiated cells have to undergo either mitosis or meiosis for tissue increase before being analyzed. The nuclei with highly fragmented chromosomes rarely survive the subsequent cell divisions. Thus, the increase in radiation dosage beyond certain level leads to a decrease in seed germination, plant survival, and vigor (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2006). All the RH studies in plants, except for studies in barley tissue culture and protoplast fusion by Wardrop et al. (2002, 2004), indicate high marker retention frequencies (75–97 %) as compared with studies conducted in animals (20–30 % retention) (Riera-Lizarazu et al. 2000, 2010; Hossain et al. 2004b; Gao et al. 2006; Kalavacharla et al. 2006; Faraut et al. 2009). The RH panel for maize chromosome 9 (Riera-Lizarazu et al. 2000) showed an average maize marker retention frequency of 85, 83, and 75 % for the 300-, 400-, and 500-Gy γ -rays treatments, respectively. Another study involving the RH panel for chromosome 1D of wheat showed an average retention of 74 % (Kalavacharla et al. 2006; Michalack de Jimenez et al. 2013). However, most of the RH mapping studies focused on generation of viable plants showed marker retention of > 90 % (Gao et al. 2004, 2006; Riera-Lizarazu et al. 2010; Yamano et al. 2010; Kumar et al. 2012a, b; Bassi et al. 2013). High marker retention would require the analysis of a larger population and identification of lines with more deletions (or breaks). This is not always possible, thus selection of RH lines with a certain range of marker loss, based on screening with few markers spread throughout the genome or region of interest can prove to be an effective strategy in reducing population size (Jones 1996; Kumar et al. 2012b; Bassi et al. 2013).

12.3.4 RH Mapping Panel Development in Plants

Strategy to develop a mapping panel depends on the objectives of a study. RH mapping method developed by Goss and Harris (1975) allows dissection of only one chromosome at a time. For example, Cox et al. (1990) used a Chinese hamster-human somatic cell hybrid line (CHG3) containing hamster chromosomes and a single copy of human chromosome 21 to irradiate at 80 Gy of X-rays, which fragmented the human chromosome 21. The fragmented donor cells were then rescued by fusing with non-irradiated hamster recipient cells (GM459). This resulted in the isolation of hybrid cell lines with various human 21 sub-chromosome fragments. These sub-chromosomal stocks were subsequently utilized for marker mapping (Cox et al. 1990). Walter et al. (1994) improved the method of Goss and Harris (1975) to dissect the entire genome and termed the process whole-genome RH (WGRH) mapping. They used human fibroblast cell lines irradiated at 30 Gy that were then fused with hamster cells to rescue fragments of all chromosomes. The resulting hybrid cell lines were used for mapping the entire human genome. We describe in the following sections the progress in developing single chromosome and whole genome mapping panels in plants.

12.3.4.1 Single Chromosome Panels

In plants, the strategy of generating single chromosome RH panels was followed in maize (Riera-Lizarazu et al. 2000; Kynast et al. 2004) and wheat (Hossain et al. 2004b; Kalavacharla et al. 2006; Paux et al. 2008; Yamano et al. 2010; Kumar et al. 2012a). The availability of maize chromosome addition lines in oat (Riera-Lizarazu et al. 1996; Ananiev et al. 1997; Kynast et al. 2001a, b) served as an excellent source for mapping of individual maize chromosomes using RH approach. Disomic maize chromosome addition lines of oat were isolated from partially self-fertile oat-maize hybrids (Riera-Lizarazu et al. 1996). Riera-Lizarazu et al. (2000) used oat-maize chromosome 9 monosomic addition line, produced by backcrossing oat-maize chromosome 9 disomic addition lines to their parental oat lines, to develop a RH panel for maize chromosome 9. The BC₁ seeds (monosomic addition of maize chromosomes) were irradiated with various doses of γ -rays to induce breaks in the added maize chromosome 9 (RH₀ generation). The irradiated (RH₀) BC₁ seeds were planted and the plants were self-pollinated to generate BC₁F₂ (RH₁) seeds. These seeds were again planted to generate viable and fertile BC₁F₂ (RH₁) plants. These fertile BC₁F₂ plants were subjected to screening for the presence or absence of maize DNA. Plants with maize chromatin among these BC₁F₂s were called RH lines and constituted a panel for maize chromosome 9. Following the same methodology, such panels are now available for most maize chromosomes (Kynast et al. 2002, 2004).

In wheat, several aneuploid stocks, such as substitution lines (Joppa and Williams 1977, 1983) and nullisomic-tetrasomic lines (Sears 1966), are available which serve as excellent material for developing RH lines for a single chromosome (Hossain et al. 2004b; Kalavacharla et al. 2006; Paux et al. 2008; Yamano et al. 2010; Kumar et al. 2012a; Bassi et al. 2013; Michalack de Jimenez et al. 2013). Hossain et al. (2004b) used an alloplasmic durum wheat line carrying heteromorphic chromosome 1D.1AL (in which a small portion of the long arm of chromosome 1D was replaced by its homoeologous counterpart from chromosome 1A) (Hossain et al. 2004c) to generate a RH panel for wheat chromosome 1D. Crosses of the male-sterile, hemizygous lines with cultivated durum produced plump viable seeds with the locus of interest and shriveled inviable seeds without the gene. One hundred plump seeds were irradiated at 350 Gy γ -rays. The plants (RH₀) from these treated seeds were crossed again with euplasmic durum line. Eighty-seven plants (RH₁) derived from plump seeds represented the RH panel for chromosome 1D (Hossain et al. 2004c). In another study, the (RH₀) plants from the irradiated durum wheat (Langdon, LDN) were crossed with durum D-genome chromosome substitution line 3D for 3B [LDN 3D (3B)] to generate a RH panel for chromosome 3B (Paux et al. 2008; Kumar et al. 2012a; Bassi et al. 2013). Similarly, single chromosome panels for wheat chromosomes 1B, 4A and 7B have been developed and characterized and more RH panels for individual chromosomes of wheat and barley are being developed by our group or in collaboration with others.

12.3.4.2 Whole Genome Panels

In plants, the strategy of Walter et al. (1994) has been used to develop WGRH panels in barley (Wardrop et al. 2002, 2004), cotton (Gao et al. 2004, 2006), wheat (Zhou et al. 2006; Riera-Lizarazu et al. 2010; Kumar et al. 2012b), and citrus (de Bona et al. 2009). Although all these studies resulted in development of WGRH panels, Wardrop et al. (2002, 2004), Zhou et al. (2006), and de Bona et al. (2009) used *in vitro* techniques to develop their panels while Gao et al. (2004, 2006), Riera-Lizarazu et al. (2010), and Kumar et al. (2012b) generated panels using *in vivo* techniques.

In barley (*H. vulgare*), 50 Gy X-ray irradiated transgenic protoplasts harboring the *bar* transgene, acting as a selectable marker (confers resistance to bialaphos), were hybridized with tobacco (*Nicotiana tabacum*) protoplasts through electrofusion. After fusion, putative RH calli were selected on the basis of exhibiting resistance to herbicide (Wardrop et al. 2002, 2004). Further confirmation of the hybrid status of these putative lines was done using the PCR amplification of a 421-base pair (bp) long fragment of the transgene. Forty of the 200 calli tested were confirmed and constituted the WGRH panel (Wardrop et al. 2002). The same strategy was further improved to obtain more DNA from hybrid calli (Wardrop 2004). In wheat, Zhou et al. (2006) described the development of WGRH by fusing the UV-irradiated protoplasts of *Bupleurum scorzonerifolium* with the protoplasts of common wheat (*T. aestivum* L.). DNA marker analysis of the hybrids demonstrated that wheat DNA was integrated into the nuclear genomes of *B. scorzonerifolium*. Following the same techniques, a WGRH panel for citrus (de Bona et al. 2009) was also developed by irradiating the microprotoplasts of *Swinglea glutinosa*, a distant relative of citrus, at various gamma ray dosages (50, 70, 100, or 200 Gy) and fusing with cv. 'Ruby Red' grapefruit or *Murcott tangor* protoplasts. AFLP analysis was then used to select for the hybrid cell lines.

In addition to these *in vitro* WGRH panels, few WGRH panels involving seed or pollen irradiation have also been reported (Gao et al. 2004, 2006; Riera-Lizarazu et al. 2010; Kumar et al. 2012b). In cotton, the irradiated pollens of *Gossypium hirsutum* L. were used to pollinate *G. barbadense* (Gao et al. 2004), and irradiated pollens of *G. barbadense* were used to pollinate *G. hirsutum* (Gao et al. 2006), to develop WGRH panel for *G. hirsutum* and *G. barbadense*, respectively. The F₁ plants carried deletions on the *G. hirsutum* or *G. barbadense* genomes. The WGRH panel developed for cotton involved interspecific hybridization and since both species had the same genomes (AD), the genotyping was limited to the markers polymorphic between the two species.

Riera-Lizarazu et al. (2010) and Kumar et al. (2012b) used mature plants obtained from irradiated seeds of hexaploid wheat (AABBDD) as a pollen source for crosses with durum wheat (AABB) plants. The resulting RH lines were quasi-pentaploids (AABBDD) with a single copy of the D-genome carrying deletions, which were characterized using genome specific markers. Two WGRH *in vivo* panels were generated for the D-genome; one for the D-genome of cultivated wheat cultivar 'Chinese Spring' and the second for the D-genome of wild diploid progenitor *Aegilops tauschii*.

12.3.4.3 Advantages and Disadvantages of Different Plant RH Panels

In animals, the only option for development of RH panels is *in vitro* cell line culture. However, plants have an advantage in that several types of resources and strategies can be employed in RH experiments. Both *in vitro* (Wardrop et al. 2002, 2004; Zhou et al. 2006; de Bona et al. 2009) and *in vivo* techniques (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2004, 2006; Hossain et al. 2004b; Kalavacharla et al. 2006; Paux et al. 2008; Yamano et al. 2010; Kumar et al. 2012a, b; Michalack de Jimenez et al. 2013; Bassi et al. 2013) have been used to develop RH panels in plants. However, these approaches have their own advantages and disadvantages.

Somatic cell hybridization, although extensively used for gene mapping in animal system, has found limited application in plants as it tends to be technically demanding and not readily available in all species. The protoplast fusion approach (Wardrop et al. 2002, 2004; Zhou et al. 2006) is laborious and finding an appropriate recipient cell line to rescue irradiation-fragmented donor chromosomes can be difficult. As plant cells can be unstable in tissue culture, maintaining a genetically stable hybrid cell line could also be problematic (Song et al. 2000; Wardrop et al. 2002). Additionally, the limited amount of DNA (Wardrop et al. 2002) and the inability to generate a viable RH plant for functional analysis are other disadvantages of the somatic hybridization technique. However, if the purpose of a study is to develop physical maps or marker scaffolds, the RH panels developed through somatic hybridization may prove useful as they typically yield higher deletion frequency (Wardrop et al. 2002, 2004) in comparison with *in vivo* panels (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2004, 2006; Hossain et al. 2004b; Kalavacharla et al. 2006; Yamano et al. 2010).

The RH population developed through seed or pollen irradiation (Riera-Lizarazu et al. 2000, 2010; Gao et al. 2004, 2006; Kynast et al. 2004; Kumar et al. 2012b; Michalack de Jimenez et al. 2013; Bassi et al. 2013) generate stable, viable, and fertile segregants. This method allows the irradiated material to pass through meiosis before being analyzed. As such the nuclei with highly fragmented chromosomes that do not survive subsequent cell divisions will be eliminated. Therefore, populations developed through seed or pollen irradiation can be used to identify genomic regions associated with a phenotype of interest (Kynast et al. 2002, 2004; Hossain et al. 2004b; Michalack de Jimenez et al. 2013; Bassi et al. 2013). The irradiation of pollen and subsequent recovery using wide crosses as done by Gao et al. (2004, 2006) in cotton, also offer an excellent opportunity to develop WGRH in plant species where suitable cytogenetic stocks are not available. *In vivo* RH panels are an ideal alternative for those species that are recalcitrant to tissue culture manipulations. Relative ease in the distribution of all or selected members of the RH panel, simplicity of long-term maintenance, and availability of unlimited resources for follow-up studies are some additional benefits of using seed/pollen for irradiation. At the same time, they can also be used for functional genomics studies (Kynast et al. 2002, 2004; Hossain et al. 2004b; Michalack de Jimenez et al. 2013; Bassi et al. 2013).

Table 12.6 List of plant species with available cytogenetic stocks suitable for developing RH panels

Donor species	Recipient species	Cytogenetic stock	Reference
<i>Aegilops speltoides</i>	<i>Triticum aestivum</i>	Addition lines	Friebe et al. 2000
<i>Allium cepa</i>	<i>Allium fistulosum</i>	Addition lines	Shigyo et al. 1996
<i>Brassica oleracea</i>	<i>B. campestris</i>	Addition lines	Quiros et al. 1987; McGrath et al. 1990
<i>Hordeum vulgare</i>	<i>Triticum aestivum</i>	Addition lines, substitution lines	Islam and Shepherd 1981; Ya-Ping et al. 2003; Szakács and Molnár-Láng 2007, 2010
<i>Beta corolliflora</i>	<i>Beta vulgaris</i>	Addition lines	Gao et al. 2001
<i>Beta patellaris</i>	<i>Beta vulgaris</i>	Addition lines	Mesbah et al. 1997
<i>Beta procumbens</i>	<i>Beta vulgaris</i>	Addition lines	van Geyt et al. 1988
<i>Beta webbiana</i>	<i>Beta vulgaris</i>	Addition lines	Reamon-Ramos and Wricke 1992
<i>Dasypyrum breviaristatum</i>	<i>T. aestivum</i>	Addition lines	Yang et al. 2008
<i>Lycopersicon esculentum</i>	<i>Solanum tuberosum</i>	Addition lines	Ali et al. 2001
<i>Oryza officinalis</i>	<i>Oryza sativa</i>	Addition lines	Jena and Khush 1986
<i>Secale cereale</i>	<i>Triticum aestivum</i>	Substitution lines	Silkova et al. 2006
<i>Solanum lycopersicoides</i>	<i>Lycopersicon esculentum</i>	Addition lines	Chetelat et al. 1998
<i>Wheat</i>	<i>Various species</i>	Addition lines, nullisomic-tetrasomic lines, substitution lines	Sears 1966, Joppa and Williams 1977, 1983
<i>Zea mays</i>	<i>Avena sativa</i>	Addition lines	Ananiev et al. 1997; Kynast et al. 2001a, b
<i>Zea mays</i>	–	Translocation lines	Sheridan and Auger 2006

12.3.5 Potential of Developing RH Panels for Any Plant Species

The few reported studies in plants clearly demonstrate that a wide range of available resources can be used for the development of RH panels. Cytogenetic stocks for various plant species offers an excellent tool for developing RH panels. In any species where the cytogenetic stocks such as addition lines, substitution lines, and introgression lines are available, they can be used for development of the RH panels for a genome, specific chromosome, or segments of a chromosome of interest. Such cytogenetic stocks are available for many species (Table 12.6). Availability of the cytogenetic stocks can be exploited in a similar fashion as done by Riera-Lizarazu et al. (2000) in maize, and by Hossain et al. (2004b) and Paux et al. (2008) in wheat to develop RH panels for any chromosome/segment of interest.

Whole genome RH panels may be generated for any plant species where sexual hybridization with a related species is possible. For example, sexual hybridization between *Corchorus olitorius* (the tossa jute) and *C. capsularis* (the white jute), two fiber-yielding cultivated species, is possible (Mia and Shaikh 1967). The high level of polymorphism (98 %) between these two species (Mir et al. 2009) makes jute a

perfect candidate to develop RH panels for genome mapping and functional studies as they also show a high level of diversity for several important phenotypes (Mir et al. 2008, 2009). The methodology followed by Gao et al. (2004, 2006), Riera-Lizarazu et al. (2010), and Kumar et al. (2012b) to develop WGRH panels can also be used in any plant species where interspecific hybridization is possible. However, due to the presence of various reproductive barriers, hybridization is restricted to sexually compatible species. Under such a scenario, somatic hybridization by combining cells of different plants may prove helpful in developing WGRH panels. Somatic hybrids are developed through protoplast fusion, thus bypassing pre- post-zygotic sexual incompatibilities and enabling the transfer of genomes between different species. With the advances in tissue culture methods, somatic hybrids have been developed for several crop plants (Fahleson and Glimelius 1999; Xiang et al. 2003; Li et al. 2004; Zhang et al. 2008; Taski-Ajdukovic et al. 2010; Wang et al. 2011; Zhang et al. 2011) offering the opportunity to develop *in vitro* WGRH panels as demonstrated by Wardrop et al. (2002, 2004).

12.3.6 Applications and Prospects for Radiation Hybrids in Plants

The implementation of RH mapping offers bright prospects for future research in plants. The utility of this approach may vary from one system to another depending upon the objectives and the availability of plant material for generating RH panels.

12.3.6.1 Radiation Hybrids for Mapping and Cloning Genes

Significance of identifying the specific location of a gene has increased with the advent of modern tools and technologies. Genomic location is essential for any comparative mapping study and/or positional cloning project. It also determines the successful utilization of genes in a breeding program using modern tools for phenotypic improvement. Cytogenetic stocks for some plant species are available to localize genes to a region on a chromosome (Table 12.6). These stocks allow the study of genes/phenotypes associated with alien/unpaired chromosome and use of a presence/absence marker detection system without the need for polymorphism. This type of material have been used in several studies such as oat-maize addition lines (Ananiev et al. 1997; Kynast et al. 2001a, b) to map ~ 350 previously unmapped EST and STS loci to maize chromosomes (Okagaki et al. 2001).

In wheat, using the property of gametocidal genes to induce chromosome breaks, a set of 436 terminal chromosome deletions were identified, which divided the whole wheat genome into smaller segments of about 40 Mb on average (Endo and Gill 1996). These terminal deletion lines have been used in a similar manner to determine relative location for many genes on chromosome segments. These deletion stocks along with nullisomic-tetrasomic (NT) lines (Sears 1954, 1966), and ditelosomic (DT) lines (Sears and Sears 1978), were used to map over 7,000 ESTs (Conley et al. 2004; Hossain et al. 2004a; Linkiewicz et al. 2004; Miftahudin et al. 2004; Munkvold et al.

2004; Peng et al. 2004; Qi et al. 2004; Randhawa et al. 2004) in wheat. Although, the use of these stocks eliminates the need for marker polymorphism, gametocidal genes are available only from certain *Aegilops* species, and transfer of this gene into lines of interest can be difficult. Moreover, the breaks induced by this system are often non-random and terminal in nature (Sakai et al. 2009). More importantly, due to the availability of a relatively small number of deletion lines, this approach offers an even lower resolution than genetic mapping. For example, several studies in wheat (Conley et al. 2004; Hossain et al. 2004a; Linkiewicz et al. 2004; Miftahudin et al. 2004; Munkvold et al. 2004; Peng et al. 2004; Qi et al. 2004; Randhawa et al. 2004) used 101 deletion lines with 119 chromosome-segment deletions for mapping. With a genome size of 16,000 Mb, this approach was able to map ESTs to chromosomal segments averaging 143.5 Mb in size. The cytogenetic materials discussed above are appropriate for mapping genes and sequences to chromosomes or a large segment of a chromosome; however, they do not allow ordering sequences on particular chromosome segments and thus are not able to provide positional information.

Radiation hybrids can be used for mapping genes/phenotypes in a similar fashion as the cytogenetic stocks, but at a much higher resolution (see Fig. 12.2). RH appears to induce random breaks including interstitial breaks (Hossain et al. 2004b; Kalavacharla et al. 2006; Paux et al. 2008; Sakai et al. 2009) and the number of breaks created per chromosome may be controlled through radiation dosage allowing for an increase in the number of genes or phenotypes studied in an experiment. A successful example of this approach was the mapping of *species cytoplasm-specific* (*scs*) gene in durum wheat (Hossain et al. 2004b). This gene was identified in an alloplasmic durum line [(1o) durum] with an introgression of a portion of chromosome 1D from *T. aestivum*. The chromosome 1D of this line segregates as a whole without recombination, making it impossible to use genetic mapping. A radiation hybrid mapping population of 87 lines allowed localization of *scs^{ae}* and eight linked markers to a region of approximately 20 Mb on the long arm of chromosome 1D. Further studies with 188 RH lines of chromosome 1D, in combination with increased markers saturation and a strategic phenotyping based on test-crossing, refined the locus position to a 1.1 Mb interval containing eight candidate genes (Michalack de Jimenez et al. 2013)

Positional cloning using a genetic mapping approach is a time consuming and mostly unsuccessful procedure similar to “reaching to the top of Mount Everest” as pointed out by Peters et al. (2003). Several major problems arise when using recombination for map-based cloning, including the need of developing a homogenous and large population, the availability of polymorphic markers, and the poor conversion of genetic to physical distances (Salvi and Tuberosa 2005). Most of these limitations can be overcome employing the RH approach.

Ideally, for fine mapping or cloning a QTL, a population developed from isogenic lines (NILs) is used, which only segregate for the region of interest. Developing such population is a time-consuming process that takes 4–6 generations. In contrast, using the RH methodology, a homogeneous population can be obtained in only one generation. In this sense, a successful example of use of small RH populations was the mapping of the *scs* gene (Hossain et al. 2004b; Michalack de Jimenez et al. 2013).

Another limitation in positional cloning is the requirement of a large number of polymorphic markers. In *Arabidopsis*, it is estimated that a high resolution map with 12,000 markers is required in order to locate a gene to a 10 Kb interval (Peters et al. 2003). To develop such high resolution maps, numerous markers need to be tested to find polymorphisms between the parents. Identifying such markers is a tedious process in positional cloning especially in non-sequenced organisms. Use of a RH panel in positional cloning is beneficial as all markers regardless of their polymorphic nature between the parents can be used. This characteristic also makes the RH method suitable for adaptation to automated, high-throughput genotyping systems. Further, it can be extrapolated that RH can also be applied to the mapping, cloning, and characterization of monomorphic genes lacking any natural variation, such as meiotic genes involved in synapsis and pairing and other loci controlling life-dependent functions.

In this sense, a good example is provided by the use of RH for wheat chromosome 3B to pinpoint a gene controlling chromosome pairing and desynapsis (*Tdes2*). This gene was initially identified on the long arm of chromosome 3B in 1954 (Sears 1954) and then reported by many other groups (see Bassi et al. 2013 for review). However, the lack of natural variation for this locus had prevented its further characterization until the advent of RH mapping. Employing 696 RH lines and a custom high-throughput platform of 140 markers, its location was refined to a 1.4 Mb interval (Bassi et al. 2013). Additionally, the use of comparative genomics suggests that this gene is indeed functional in many cereals, confirming that the use of RH for the study of meiotic genes in one species could impact many other organisms.

Finally, the accurate estimation of physical distance is not possible in the case of genetic mapping approaches due to a wide variation in recombination frequency across the length of the chromosome. For instance, in *Arabidopsis*, on average 1% recombination occurs in 250 kb, while in centromeric regions the recombination rate decreases to 1% in 1,750 kb (Lukowitz et al. 2000). Also, the recombination rate is higher in closely related lines compared to distantly related lines (Peters et al. 2003). Mapping distances in RH populations are estimated based on physical breakage that occur relatively randomly in the genome. Thus the distances obtained from RH mapping are a better estimation of actual physical distances (Kumar et al. 2012a).

Radiation-induced deletion lines were used to narrow down the location of a homeologous pairing suppressor (*ph1*) locus, located on chromosome 5B of hexaploid wheat (Roberts et al. 1999; Al-Kaff et al. 2008). In order to identify deletions of the region spanning the gene of interest, various X-ray, γ -ray, as well as fast neutron mutant populations were generated. The *Ph1* locus was found in a region between two loci (*Xrgc846* and *Xpsr150A*) (Roberts et al. 1999). Recently, using deletion mutants and expression profiling, the candidates for *Ph1* locus were identified to be a *cdk*-like gene cluster (Al-Kaff et al. 2008).

In another study, researchers used two methods to delineate the region of the *Q* locus in wheat and identify candidate genes (Faris et al. 2003). The first approach was using BACs from a related diploid species that were sequenced for chromosome walking. After sequencing multiple BACs, several markers were co-localizing with the *Q* locus. The second approach was screening radiation mutants to phenotypically

identify lines with a loss of the *Q* locus. These lines were then characterized with the genetic markers from the BAC clones. One of the lines showed a breakage between the *Q* locus and a genetically mapped linked locus. Employing radiation mutants allowed to more quickly identify a candidate gene region by eliminating the need for more BAC sequencing and reducing the size of the region containing candidate genes. Radiation hybrids developed by methods describe earlier (Kalavacharla et al. 2006; Paux et al. 2008; Kumar et al. 2012a, b) offer a much higher resolution than those employed for *Ph* or *Q* locus. This can effectively facilitate mapping and positional cloning by shortening the amount of time needed for high resolution mapping and generating new mutants for proof of cloning.

12.3.6.2 RH Mapping and Forward/Reverse Genetics

Forward genetics are the approaches that start with the observable variation and explore the underlying genetic mechanism for a particular variation. On the other hand, reverse genetics is the approach that starts with the DNA sequence, which is the ultimate source of all variation and identifies the coded phenotype.

The application of radiation hybrids for reverse genetics studies has been mostly limited to insects and animals. Using *Drosophila melanogaster* X-rays induced mutations in the chaoptic (*chp*) gene sequence resulted in decreased levels for a protein leading to functional phenotypic analysis (Van Vactor et al. 1988). In order to study, the function of kinesin heavy chain (*Khc*) gene in *D. melanogaster*, γ -irradiation was used to create mutations in chromosome 2 changing the phenotype of the fly. The phenotypic changes resulted in fly developmental changes and were essential for neuromuscular system effecting growth and movement. To aid in the identification of lines of interest, a selectable marker for eye color was used to find pupa with deletions within the region of chromosome 2 of interest (Saxton et al 1991). A similar approach was used in stem cells of mice which contained a cassette to identify genes of interest at the *DI7Aus9* locus. The lines were irradiated to observe phenotypic changes due to loss of genes at the *DI7Aus9* locus (You et al. 1997). This approach created viable mutated individuals with phenotypes of runt size, and shortened or kinked tail, and a possible approach to discover functions for mouse genes.

In addition to cloning genes as described in the previous section, the forward genetic approach using radiation induced mutations has been applied in *Arabidopsis*, rice, tomato, and wheat. The use of radiation in *Arabipodsis* and rice resulted in monogenic recessive alleles induced by radiation (Koornneef et al. 1982). For tomato, about 4,000 mutant lines are available for the estimated 30,000 genes. These represent an opportunity to relate phenotype to about 1/6 of the genes in the tomato genome. However, for the remaining portion of genes, it may be possible to discover their role by changes in radiation dosage to create additional or larger mutations. About 4 % of the tomato radiation lines showed a phenotype related to changes in leaf, fruit, or flower morphology (Matsukura et al. 2007). It is anticipated that these materials will help in understanding the genes involved in sugar content of fruit, important to the tomato industry (Matsukura et al. 2007). Work on *Arabidopsis* in the early 1980s also

identified a number of phenotypes resulting from radiation of the genome. In an M_2 population, about 0.6 % mutants were created by fast neutron radiation. Phenotypic mutants reported include gibberillic acid sensitive, hypocotyl morphology, narrow leaves, double flower, and trichome mutants (Koornneef et al. 1980, 1982). These mutants led to the cloning of genes involved in testa color such as chalcone flavanone isomerase (*CHI*) and dihydroflavonol 4-reductase (*DFR*) (Shirley et al. 1992), and the placement of genes related to plant height and flowering onto chromosomes (Koornneef et al. 1991). Recently this approach has also been applied to rice with phenotypes for leaf color and growth habit, heading date, plant height, and lethality observed in radiation mutants. So far, deletion rate in available plant RH panels have been lower than those observed in animals. Low deletion rate in RH panels make it possible to develop *in vivo* viable populations. Compared to *in vitro* RH hybrid panels in animals, plant RHs are easy to develop, reproducible, and are ethically acceptable to work with. Such populations are great candidates in forward/reverse genetic studies to associate changes in gene sequence with phenotypes and vice versa in positional cloning of genes underlying mutant phenotypes. Overall, RH panels could effectively facilitate positional cloning by inducing new mutants and high resolution mapping in a short period of time. It is also important to consider that radiation might cause DNA rearrangements other than deletion that also affect phenotypes. In this case map location of loci involved in rearrangement would be a distortion of the original location. There is a small possibility that genes underlying phenotype of interest are involved with one or both breakpoints of rearrangement (Lukowitz et al. 2000).

12.3.6.3 Radiation Hybrid Mapping and Comparative Genomics

Comparative genomic studies across species can shed light on chromosome evolution as well as provide practical information for understanding genome structure. Many synteny studies in animals have been facilitated by utilizing RH maps and sequences of model species. After the completion of human genome sequence, orthologous regions in other closely related species like monkeys, apes, chimpanzees and gorillas, were identified and the RH mapping was the approach which enabled comparative studies between them. For example, using 93 hybrid cell lines and 84 loci, a comparative map of chromosomes 7 and 9 of *Macaca mulatta* and chromosomes 14 and 15 of human was generated. This study revealed a complex evolutionary scenario which involved a loss of the interstitial centromere in the two ancestral chromosomes (14 and 15) followed by the evolution of two novel centromeres at the end of both chromosomes (Murphy et al. 2001). Another study reported a high-resolution RH map of ovine chromosome segments homologous to human chromosome 6 (HSA6). This map was constructed with 251 markers using a whole-genome RH panel. Syntenic regions between HSA6 and three ovine chromosomes (OAR8, 9, and 20) were observed. Two of these syntenic regions were previously reported in chromosome painting studies. In addition, a novel homology of a small centromeric region on the OAR9 and HSA6 was also reported (Wu et al. 2008). This illustrates the power of

RH mapping in detecting small syntenic regions within chromosomes of different species which can provide essential information regarding the evolution of various species.

A large number of studies using whole genome comparative analysis have been done in mammals to explore chromosome evolution. This was possible not only because of the availability of the complete genome sequences, but also due to the existence of a well-established RH mapping procedure. Murphy et al. (2005) integrated the information from the RH mapping of cat, cattle, dog, pig, and horse with the information of the whole genome sequences of human, mouse, and rat to address fundamental questions related to mammalian chromosome evolution. These authors identified 1,159 pairwise homologous synteny blocks (HSBs) between the genomes of human and the six other species and constructed an evolutionary scenario depicting the rearrangements between all the genomes and their ancestors.

The whole genome comparative analysis also gives important information about the physical clustering of gene families across species and also helps in identifying structural linkage associations preserved for millions of years across thousands of species (O'Brien et al. 1999). These preserved blocks could be considered as "frozen accidents" or selectively retained due to their association with particular developmental or biochemical pathways in a cell.

Compared to animals, there are only few studies related to the chromosomal evolution in grasses and they are mostly restricted to comparative genome analysis of a few species that are either fully sequenced (Vogel et al. 2010) or for which some information is available from the low resolution genetic and the chromosomal bin maps (Salse et al. 2008; Luo et al. 2009) which limits our understanding of the precise events and mechanisms in the evolution of a species. Although only a few plant species have been fully sequenced, the amount of sequence information available for many species is rapidly growing due to the development of the next generation sequencing technologies, and currently there are several ongoing sequencing projects for various plant species (Feuillet et al. 2011). Consequently, sequenced model plant species became a very important tool in comparative studies (Paterson et al. 2000). *Arabidopsis thaliana* is an example of such species which became a model plant for synteny analysis in *Brassicaceae* (Paterson et al. 2000). The sequencing and assembly of *A. thaliana* genome was relatively easy due to its small size (125 Mb) and low amount of repetitive elements (about 1%) (Surzycki and Belknap 1999). Comparing gene-based maps of a species of interest with a sequence of a model species enables the identification of large, conserved syntenic regions which in turn help in localization of important genes and understanding the evolution of genomes.

Early comparative studies were based on EST-derived RFLP maps of closely related species. Most of the genetic maps provided very low resolution of around 10 cM which allowed only detection of conserved regions at the macroscopic level (macro-colinearity; Kurata et al. 1994; Deynze et al. 1995; Sarma et al. 2000). The limited resolution is often a problem in genetic mapping which affects comparative analysis. The low resolution does not allow for separation of two or more gene-based markers making it impossible to determine whether these genes maintained their order across species or if any ancestral rearrangements occurred. Another issue

affecting the synteny analysis is a need for polymorphism in order to genetically map a given marker on the mapping population. Unfortunately, the selection pressure tends to get rid of any non-advantageous allelic polymorphism in the gene space, making the gene-based markers monomorphic among individuals of the same species. For example, in hexaploid wheat there was 9.9% polymorphism observed for EST-derived SSRs compared to 35.5% polymorphism observed in genomic SSRs (Xue et al. 2008).

Radiation hybrid mapping offers a new strategy for mapping monomorphic markers and also provides a significantly higher mapping resolution (de Pontbriand et al. 2002; Gautier and Eggen 2005) suitable for studying syntenic relationships among species even at the microscopic level (micro-colinearity). Recently, Michalak et al. (2009) studied the micro-colinearity of orthologous nuclear genes encoding mitochondrial proteins from a short segment of chromosome 10 of rice which is collinear with chromosome 1D of wheat. In this study, a panel of 94 RH lines was used to genotype 11 mitochondrial-related nuclear genes. Chromosomal rearrangements at a micro-colinearity level were observed between wheat chromosome 1D and *Brachypodium* Super contig 8 (corresponding to the chromosome Bd3) and rice chromosome 10. Interestingly, a perfect micro-colinearity between rice and *Brachypodium* was also observed. This might be related to the fact that wheat being a recent polyploid has buffering effects of polyploidy which allows for more changes than could be possible in the diploid species (Michalak et al. 2009). An extension of this work with 188 RH lines, in combination with the use of 57 EST-derived markers and an on-demand synteny tool (Alnemer et al. 2013), allowed to determine the ancestral origin of the centromere of chromosome 1D (Michalack de Jimenez et al. 2013). The resolution provided by the RH approach confuted a theory previously validated by recombination-based mapping, and instead supported a neo-origin of the wheat centromere as result of paleofusion of two ancestral chromosomes (Michalack de Jimenez et al. 2013).

Development of RH maps has contributed extensively to comparative analysis and evolutionary studies in animal systems. However, in plants, the use of RHs has been limited which may be due to the fact that genetic populations were easy to develop and till now, the need for developing higher resolution maps was not recognized. RH mapping allows the construction of high-resolution physical maps after survey sequencing at low coverage (1–2x) allowing easy assembly of important segments (Donthu et al. 2009). High-resolution RH comparative maps can provide the same information as genome sequence assemblies (which require much higher genome coverage) but at a fraction of the cost (Hitte et al. 2005). Thus, thousands of gene-based markers can be developed from low coverage sequencing of important species and used for high resolution RH-based physical maps and comparative analysis.

12.3.6.4 RH Mapping and Survey Sequencing

In last few years, several next-generation sequencing platforms have been developed, resulting in availability of vast sequence information (Gupta 2008; Metzker 2010).

The generation of the sequence data has dramatically accelerated genetic research and generated much new information. Next generation sequencing technologies have also led to the availability of genome sequences for several plant species and the projects are underway for many others (for review, see Feuillet et al. 2011). Although, there is a continuous decrease in the cost of sequencing, it still remains quite expensive to generate a 8–10x sequence coverage, typically used in the case of assembled plant genomes (Feuillet et al. 2011). A survey sequence with 1–2x genome coverage has been shown to provide a substantial amount of information (Hitte et al. 2005). Although, survey sequencing lacks long-range continuity required for whole genome assembly, it provides valuable genomic resources for mapping studies and if a reference sequence of related species is available, comparative analysis can reveal conserved gene order between different genomes. Survey sequencing when combined with radiation hybrid mapping can prove valuable in extracting genomic information at a lower cost. In canines, 90 % of the over 10,000 gene markers derived from 1.5x survey sequencing (Kirkness et al. 2003) were located within 18,201 genes derived from automated annotation of the 7.5x assembly (Hitte et al. 2005). Based on this study in canines, it was suggested that RH mapping combined with 0.5–1x survey sequence information is capable of generating a high-quality comparative map resource.

In large genome plant species where a high amount of repetitive DNA and ploidy levels make sequencing more expensive and assembly computationally challenging, sequencing of the low-copy, non-repetitive targets or gene space using transcriptome sequencing (Barbazuk et al. 2007; Novaes et al. 2008) or reduced representation genomic libraries (Van Orsouw et al. 2007) by next-generation sequencing technologies can substantially reduce the cost. Development of thousands of molecular markers from these genic sequences and high resolution radiation hybrid based physical mapping will reveal the organization of gene space and provide an excellent resource for comparative analysis and gene cloning, in addition to serving as a marker scaffold for genome sequencing projects. RH mapping and sequence information in plants will open a new window for genomic studies in the near future.

Acknowledgments This work was supported by funding from the National Science Foundation, Plant Genome Research Program (NSF-PGRP) grant No. DBI-0822100 to SFK.

References

- Ahloowalia BS, Maluszynski M, Nichterlein K (2004) Global impact of mutation-derived varieties. *Euphytica* 135:187–204
- Ahmed EA, Philippens MEP, Kal HB et al (2010) Genetic probing of homologous recombination and non-homologous end joining during meiotic prophase in irradiated mouse spermatocytes. *Mutat Res* 688:12–18
- Akhunov ED, Goodyear AW, Geng S et al (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res* 5:753–763

- Ali SNH, Ramanna MS, Jacobsen E, Visser RGF (2001) Establishment of a complete series of a monosomic tomato chromosome addition lines in the cultivated potato using RFLP and GISH analyses. *Theor Appl Genet* 103:687–695
- Al-Kaff N, Knight E, Bertin I et al (2008) Detailed dissection of the chromosomal region containing the *Ph1* locus in wheat *Triticum aestivum*: with deletion mutants and expression profiling. *Ann Bot* 105:6075–6080
- Alnemer LM, Seetan RI, Bassi FM, et al (2013) Wheat Zapper: a flexible online tool for colinearity studies in plants. *Funct Integr Genomics* 13:11–17
- Ananiev EV, Riera-lizarazu O, Rines HW, Phillips RL (1997) Oat-maize chromosome addition lines: a new system for mapping the maize genome. *Proc Natl Acad Sci U S A* 94: 3524–3529
- Anderson EG, Longley AE, Li CH, Retherford KL (1949) Hereditary effects produced in maize by radiations from the Bikini atomic bomb I. Studies on seedlings and pollen of the exposed generation. *Genetics* 34:639–646
- Argonne National Laboratory (2005) Ionizing radiation. Human health fact sheet, August
- Barbazuk WB, Emrich SJ, Chen HD et al (2007) SNP discovery via 454 transcriptome sequencing. *Plant J* 51:910–918
- Bassi FM, Kumar A, Zhang Q et al (2013) Radiation hybrid QTL mapping of *Tdes2* involved in the first meiotic division of wheat. *Theor Appl Genet* 126(8):1977–1990
- Bertagne-Sagnard B, Fouilloux G, Chupeau Y (1996) Induced albino mutations as a tool for genetic analysis and cell biology in flax (*Linum usitatissimum*). *J Exp Bot* 47:189–194
- Blakeslee AF (1935) Obituary: Hugo De Vries: 1848–1935. *Science* 81:581–582
- Blonstein AD, Parry AD, Horgan R, King PJ (1991a) A cytokinin-resistant mutant of *Nicotiana plumbaginifolia* is wilted. *Planta* 183:244–250
- Blonstein AD, Stirnberg P, King PJ (1991b) Mutants of *Nicotiana plumbaginifolia* with specific resistance to auxin. *Mol Gen Genetics* 228:361–371
- Britt AB (1996) DNA damage and repair in plants. *Annu Rev Plant Physiol Plant Mol Biol* 47:75–100
- Britt AB (1999) Molecular genetics of DNA repair in higher plants. *Trends Biotechnol* 4:20–25
- Carroll BJ, Gresshoff PM, Delves AC (1988) Inheritance of supermodulation in soybean and estimation of the genetically effective cell number. *Theor Appl Genet* 76:54–58
- Chetelat RT, Rick CM, Cisneros P et al (1998) Identification, transmission, and cytological behavior of *Solanum lycopersicon* Dun. monosomic alien addition lines in tomato (*Lycopersicon esculentum* Mill.). *Genome* 41:40–50
- Conley E, Nduati V, Gonzalez-Hernandez JL et al (2004) A 2,600-locus chromosome bin map of wheat group 2 reveals interstitial gene-rich islands and colinearity with rice. *Genetics* 168:625–637
- Cox DR, Burmeister M, Price ER et al (1990) Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. *Science* 250:245–250
- de Bona CM, Stelly D, Miller JC, Louzada ES (2009) Fusion of protoplasts with irradiated microprotoplasts as a tool for radiation hybrid panel in citrus. *Pesq Agropec Bras* 44:1616–1623
- de Pontbriand A, Wang XP, Cavaloc Y et al (2002) Synteny comparison between apes and human using fine-mapping of the genome. *Genomics* 80:395–401
- Deloukas P, Schuler GD, Gyapay G et al (1998) A physical map of 30,000 human genes. *Science* 282:744–746
- Deynze AE, Nelson JC, Yglesias ES et al (1995) Comparative mapping in grasses. Wheat relationships. *Mol Gen Genetics* 248:744–754
- Donthu R, Lewin HA, Larkin DM (2009) SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence. *BMC Res Notes* 2:148
- Driscoll CJ, Jensen NF (1963) A genetic method for detecting induced intergeneric translocations. *Genetics* 48:459–468
- Elsik CG, Tellam RL et al (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324:522–528
- Endo TR, Gill BS (1996) The deletion stocks of common wheat. *Heredity* 87:295–307

- Erayman M, Sanduh D, Sidhu D et al (2004) Demarcating the gene-rich regions of the wheat genome. *Nucleic Acids Res* 32:3546–3565
- Fahleson J, Glimelius K (1999) Protoplast fusion for symmetric somatic hybrid production in Brassicaceae. *Methods Mol Biol* 111:195–209
- Faraut T, de Givry S, Hitte C et al (2009) Contribution of radiation hybrids to genome mapping in domestic animals. *Cytogenet Genome Res* 126:21–33
- Faris JD, Fellers JP, Brooks SA, Gill BS (2003) A bacterial artificial chromosome contig spanning the major domestication locus *Q* in wheat and identification of a candidate gene. *Genetics* 164:311–321
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trends in Plant Sci* 16:77–88
- Freisleben R, Lein A (1944) Rontgeninduzierte Mutationen bei Gerste. *Zuchter* 16:49–64
- Friebe B, Qi LL, Nasuda S et al (2000) Development of a complete set of *Triticum aestivum*-*Aegilops speltoides* chromosome addition lines. *Theor Appl Genet* 101:51–58
- Gao D, Guo D, Jung C (2001) Monosomic addition lines of *Beta corolliflora* in sugar beet: cytological and molecular marker analysis. *Theor Appl Genet* 103:240–247
- Gao W, Chen ZJ, Yu JZ et al (2004) Wide-cross whole-genome radiation hybrid mapping of cotton (*Gossypium hirsutum* L.). *Genetics* 167:1317–1329
- Gao W, Chen ZJ, Yu JZ et al (2006) Wide-cross whole-genome radiation hybrid mapping of the cotton (*Gossypium barbadense* L.) genome. *Mol Gen Genomics* 275:105–113
- Gautier M, Eggen A (2005) The construction and use of radiation hybrid maps in genomic research. *Encyclopedia of genetics, genomics, proteomics and bioinformatics*
- Gibbs RA, Weinstock GM, Metzker ML et al (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 428:493–521
- Goss SJ, Harris H (1975) New method for mapping genes in human chromosomes. *Nature* 255:680–684
- Gottschalk W (1989) *Allgemeine Genetik*. 3 Aufl. Georg Thieme Verlag, Stuttgart
- Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* 26:602–611
- Gustafsson A (1954) Swedish mutation work in plants: background and present organization. *Acta Agriculturae Scandinavica IV* 3:361–364
- Gyapay G, Schmitt K, Fizames C et al (1996) A radiation hybrid map of the human genome. *Hum Mol Genet* 5:339–346
- Harten AM van (1998) *Mutation breeding: theory and practical applications*. Cambridge University Press, London
- Hitte C, Madeoy J, Kirkness EF et al (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat Rev Genet* 6:643–648
- Hlatky L, Sachs RK, Vazquez M, Cornforth MN (2002) Radiation-induced chromosome aberrations: insight gained from biophysical modeling. *Bioessays* 24:714–723
- Hodgdon AL, Marcus AH, Arenaz P et al (1981) Ontogeny of the barley plant as related to mutation expression and detection of pollen mutations. *Environmental Health Perspectives* 37:5–7
- Hoffmann W, Zoschke U (1955) Rontgenmutationen beim Flachs (*Linum usitatissimum* L.) *Zuchter* 25:199–206
- Hossain KG, Kalavacharla V, Lazo GR et al (2004a) A chromosome bin map of 2,148 EST loci of wheat homoeologous group 7. *Genetics* 168:687–699
- Hossain KG, Riera-lizarazu O, Kalavacharla V et al (2004b) Radiation hybrid mapping of the species cytoplasm-specific (*scs^{ae}*) gene in wheat. *Genetics* 168:415–423
- Hossain KG, Riera-lizarazu O, Kalavacharla V et al (2004c) Molecular cytogenetic characterization of an alloplasmic durum wheat line with a portion of chromosome 1D of *Triticum aestivum* carrying the *scs^{ae}* gene. *Genome* 47:206–214
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Islam AKMR, Shepherd KW (1981) Production of disomic wheat barley chromosome addition lines using *Hordeum bulbosum* crosses. *Genet Res* 37:215–219

- Jena KK, Khush GS (1986) Production of monosomic alien addition lines of *Oryza sativa* having a single chromosome of *O. officinalis*. In: Rice Genetics. IRRI, Manila, pp 199–208
- Johri MM, Coe EH (1983) Clonal analysis of corn plant development I. The development of the tassel and ear shoot. *Dev Biol* 97:154–172
- Jones HB (1996) Hybrid selection as a method of increasing mapping power for radiation hybrids. *Genome Res* 6:761–769
- Joppa LR, Williams ND (1977) D-genome substitution monosomics of durum wheat. *Crop Sci* 17:772–776
- Joppa LR, Williams ND (1983) The Langdon durum disomic-substitutions: development, characteristics, and uses. *Agron Abstr* pp 68
- Kalavacharla V, Hossain K, Gu Y et al (2006) High-resolution radiation hybrid map of wheat chromosome 1D. *Genetics* 173:1089–1099
- Karere GM, Froenicke L, Millon L et al (2008) A high-resolution radiation hybrid map of *Rhesus macaque* chromosome 5 identifies rearrangements in the genome assembly. *Genomics* 92:210–218
- Kavathas P, Bach FH, DeMars R (1980) Gamma ray-induced loss of expression of HLA and glyoxalase I alleles in lymphoblastoid cells. *Proc Natl Acad Sci U S A* 77:4251–4255
- Kirkness EF, Bafna V, Halpern AL et al (2003) The dog genome: survey sequencing and comparative analysis. *Science* 301:1898–1903
- Konzak CF (1954) Stem rust resistance in oats induced by nuclear radiation. *Agron J* 46:538–540
- Koornneef M, van der Veen JH (1980) Induction and analysis of gibberellin-sensitive mutants in *Arabidopsis thaliana* (L.) Heynh. *Theor Appl Genet* 58:257–263
- Koornneef M, Dellaert LWM, van der Veen JH (1982) EMS- and radiation-induced mutation frequencies at individual loci in *Arabidopsis thaliana* (L.) Heynh. *Mutat Res* 93:109–123
- Koornneef M, Hanhart CJ, Veen JH (1991) A genetic and physiological analysis of late flowering mutants in *Arabidopsis thaliana*. *Mol Gen Genet* 229:57–66
- Kumar A, Bassi FM, Paux E et al (2012a) DNA repair and crossing over favor similar chromosome regions as discovered in radiation hybrid of *Triticum*. *BMC Genomics* 13:339
- Kumar A, Simons K, Iqbal MJ et al (2012b) Physical mapping resources for large plant genomes: radiation hybrids for wheat D-genome progenitor *Aegilops tauschii* accession AL8/78. *BMC genomics* 13:597
- Kurata N, Moore G, Nagamura Y et al (1994) Conservation of genome structure between rice and wheat. *Nature Biotechnology* 12:276–278
- Kynast RG, Riera-Lizarazu O, Vales MI et al (2001a) A complete set of maize individual chromosome additions to the oat genome. *Plant Physiol* 125:1216–1227
- Kynast RG, Okagaki RJ, Odland WE et al (2001b) Oat-maize chromosome manipulation for the physical mapping of maize sequences. *Maize Genet Coop Newslett* 75:54–55
- Kynast RG, Okagaki RJ, Rines HW, Phillips RL (2002) Maize individualized chromosome and derived radiation hybrid lines and their use in functional genomics. *Funct Integr Genomic* 2:60–69
- Kynast RG, Okagaki RJ, Galatowitsch MW et al (2004) Dissecting the maize genome by using chromosome addition and radiation hybrid lines. *Proc Natl Acad Sci U S A* 101:9921–9926
- Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Le Signor C, Savoie V, Aubert G et al (2009) Optimizing TILLING populations for reverse genetics in *Medicago truncatula*. *Plant Biotech J* 7:430–441
- Li C, Xia G, Xiang F et al (2004) Regeneration of asymmetric somatic hybrid plants from the fusion of two types of wheat with Russian wild rye. *Plant Cell Rep* 23:461–467
- Lindblad-Toh K, Wade CM, Mikkelsen TS et al (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438:803–819
- Linkiewicz AM, Qi LL, Gill BS et al (2004) A 2,500-locus bin map of wheat homoeologous group 5 provides new insights on gene distribution and colinearity with rice. *Genetics* 168:665–676
- Lönnig WE (2005) Mutation breeding, evolution, and the law of recurrent variation. *Recent Res Devel Genet Breeding* 2:45–70

- Lukowitz W, Gilmore CS, Scheible WR (2000) Positional cloning in *Arabidopsis*. Why it feels good to have a genome initiative working for you. *Plant Physiol* 123:795–805
- Lundqvist U (1986) Svalöf 1886–1986, research and results in plant breeding. In: Olsson G (ed), Stockholm, pp 76–84
- Lunetta KL, Boehnke M, Lange K, Cox DR (1996) Selected locus and multiple panel models for radiation hybrid mapping. *Am J Hum Genet* 59:717–725
- Luo MC, Deal KR, Akhunov ED et al (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in *Triticeae*. *Proc Natl Acad Sci U S A* 106:15780–15785
- Maluszynski M, Ahloowalia BS, Sigurbjornsson B (1995) Application of *in vivo* and *in vitro* mutation techniques for crop improvement. *Euphytica* 85:303–315
- Maluszynski M, Nichterlein K, van Zanten L, Ahloowalia BS (2000) Officially released mutant varieties—the FAO/IAEA Database. *Mut Breed Rev* 12:1–84
- Matsukura C, Aoki K, Fukuda N et al (2007) Comprehensive resources for tomato functional genomics based on the miniature model tomato micro-tom. *Current Genomics* 9:436–443
- McGrath JM, Quiros CF, Harada JJ, Lanbry BS (1990) Identification of *Brassica oleracea* monosomic alien-chromosome addition lines with molecular markers reveals extensive gene duplication. *Mol Gen Genet* 223:198–204
- Mesbah M, De Bock TSM, Sandbrink JM et al (1997) Molecular and morphological characterization of monosomic additions in *Beta vulgaris*, carrying extra chromosomes of *B. procumbens* or *B. patellaris*. *Mol Breed* 3:147–157
- Metzker ML (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46
- Mézard C (2006) Meiotic recombination hotspots in plants. *Biochem Soc Trans* 34:531–534
- Mia MM, Shaikh AQ (1967) Gamma radiation and interspecific hybridization in jute. *Euphytica* 16:61–68
- Michalak MK, Ghavami F, Lazo GR et al (2009) Evolutionary relationship of nuclear genes encoding mitochondrial proteins across four grass species and *Arabidopsis thaliana*. *Maydica* 54:471–483
- Michalak de Jimenez MK, Bassi FM, Ghavami F, et al (2013) A radiation hybrid map of chromosome 1D reveals synteny conservation at a wheat speciation locus. *Funct Int Gen* 13:19–32
- Miftahudin, RK, Ma XF et al (2004) Analysis of EST loci on wheat chromosome group 4. *Genetics* 168:651–663
- Mir RR, Rustgi S, Sharma S et al (2008) A preliminary genetic analysis of fibre traits and the use of new genomic SSRs for genetic diversity in jute. *Euphytica* 161:413–427
- Mir RR, Banerjee S, Das M et al (2009) Development and characterization of large scale simple sequence repeats in jute. *Crop Sci* 49:1687–1694
- Munkvold JD, Greene RA, Bermudez-Kandianis CE et al (2004) Group 3 chromosome bin maps of wheat and their relationship to rice chromosome 1. *Genetics* 168:639–650
- Murphy WJ, Page JE, Smith C et al (2001) A radiation hybrid mapping panel for the rhesus macaque. *Heredity* 92:516–519
- Murphy WJ, Larkin DM, Everts-van der Wind A et al (2005) Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. *Science* 309:613–617
- Novaes E, Drost DR, Farmerie WG et al (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* 9:312
- O'Brien SJ, Menotti-Raymond M, Murphy WJ et al (1999) The promise of comparative genomics in mammals. *Science* 286:458–481
- Okagaki RJ, Kynast RG, Livingston SM et al (2001) Mapping maize sequences to chromosomes using oat-maize chromosome addition materials. *Plant Physiol* 125:1228–1235
- Olivier M, Aggarwal A, Allen J et al (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science* 291:1298–1302
- Oltmann W (1950) Züchterische Auswertung röntgeninduzierter Mutationen an physiologischen Merkmalen bei Winterweizen. *Zeits Pflanzenz* 29:76–89
- Page DR, Grossniklaus U (2002) The art and design of genetic screens: *Arabidopsis thaliana*. *Nat Rev Genet* 3:124–136

- Paterson AH, Bowers JE, Burow MD et al (2000) Comparative genomics of plant chromosomes. *Plant Cell* 12:1523–1539
- Paux E, Sourdille P, Salse J et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322:101–104
- Pelsy F, Kronenberger J, Pollien JM, Caboche M (1991) M₂ seed screening for nitrate reductase deficiency in *Nicotiana plumbaginifolia*. *Plant Sci* 76:109–114
- Peng JH, Zadeh H, Lazo GR et al (2004) Chromosome bin map of expressed sequence tags in homoeologous group 1 of hexaploid wheat and homoeology with rice and *Arabidopsis*. *Genetics* 168:609–623
- Peters JL, Crude F, Gerats T (2003) Forward genetics and map-based cloning approaches. *Trends Plant Sci* 8:484–491
- Qi LL, Echaliier B, Chao S et al (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 168:701–712
- Quiros CF, Ochoa O, Kianian SF, Douches D (1987) Analysis of the *Brassica oleracea* genome by the generation of *B. campestris-oleracea* chromosome addition lines: characterization by isozymes and rDNA genes. *Theor Appl Genet* 74:758–766
- Randhawa HS, Dilbirligi M, Sidhu D et al (2004) Deletion mapping of homoeologous group 6-specific wheat ESTs. *Genetics* 168:677–686
- Reamon-Ramos SM, Wricke G (1992) A full set of monosomic addition lines in *Beta vulgaris* from *Beta webbiana*: morphology and isozyme markers. *Theor Appl Genet* 84:411–418
- Riera-Lizarazu O, Rines HW, Phillips RL (1996) Cytological and molecular characterization of oat × maize partial hybrids. *Theor Appl Genet* 93:123–135
- Riera-Lizarazu O, Vales MI, Ananiev EV et al (2000) Production and characterization of maize chromosome 9 radiation hybrids derived from an oat-maize addition line. *Genetics* 156:327–339
- Riera-Lizarazu O, Leonard JM, Tiwari VK, Kianian SF (2010) A method to produce radiation hybrids for the D-genome chromosomes of wheat (*Triticum aestivum* L.). *Cytogenet Genome Res* 129:234–240
- Riley R, Law CN (1984) Chromosome manipulation in plant breeding: progress and prospects. *Stadler Genet Symp* 16:301–322
- Roberts MA, Reader SM, Dalgliesh C et al (1999) Induction and characterization of *Ph1* wheat mutants. *Genetics* 153:1909–1918
- Saintenac C, Falque M, Martin OC et al (2009) Detailed recombination studies along chromosome 3B provide new insights on crossover distribution in wheat (*Triticum aestivum* L.). *Genetics* 181:393–403
- Sakai K, Nasuda S, Sato K, Endo TR (2009) Dissection of barley chromosome 3H in common wheat and a comparison of 3H physical and genetic maps. *Genes Genet Syst* 84:25–34
- Salse J, Bolot S, Throude M et al (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 20:11–24
- Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Sci* 10:297–304.
- Sarma RN, Fish L, Gill BS, Snape JW (2000) Physical characterization of the homoeologous group 5 chromosomes of wheat in terms of rice linkage blocks, and physical mapping of some important genes. *Genome* 43:191–198
- Saxton WM, Hicks J, Goldstein LSB, Raff EC (1991) Kinesin heavy chain is essential for viability and neuromuscular functions in *Drosophila*, but mutants show no defect in mitosis. *Cell* 64:1093–1102
- Schnurbusch T, Collins NC, Eastwood RF et al (2007) Fine mapping and targeted SNP survey using rice-wheat gene colinearity in the region of the *Bo1* boron toxicity tolerance locus of bread wheat. *Theor Appl Genet* 115:451–461
- Sears ER (1954) The aneuploids of common wheat. *Missouri Agr Exp Sta Res Bull* 572:1–58
- Sears ER (1956) The transfer of leaf-rust resistance from *Aegilops umbellulata* to wheat. *Brookhaven Symp Biol* 9:1–22
- Sears ER (1966). Nullisomic-tetrasomic combinations in hexaploid wheat. In: Riley R, Lewis KR (eds) *Chromosome manipulations and plant genetics*, Oliver and Boyd Ltd., Edinburgh, p 2945

- Sears ER (1993) Use of radiation to transfer alien chromosome segments to wheat. *Crop Sci* 33:897–901
- Sears ER, Sears LMS (1978) The telocentric chromosomes of common wheat. In: Ramanujam S (ed) Proceedings of the fifth international wheat genetics symposium, New Delhi, pp 389–407
- Sengupta S, Harris CC (2005) p53: traffic cop at the crossroads of DNA repair and recombination. *Nat Rev Mol Cell Biol* 6:44–55
- Shebeski LH, Lawrence T (1954) The production of beneficial mutations in barley by irradiation. *Canad J Agri Sci* 34:1–9
- Sheridan WF, Auger DL (2006) Construction and uses of new compound B-A-A maize chromosome translocations. *Genetics* 174:1755–1765
- Shigyo M, Tashiro Y, Isshiki S, Miyazaki S (1996) Establishment of a series of alien monosomic addition lines of Japanese bunching onion (*Allium fistulosum* L) with extra chromosomes from shallot (*A. cepa* L. *Aggregatum* group). *Genes Genet Syst* 71:363–371
- Shirley BM, Hanley S, Goodman HM (1992) Effects of ionizing radiation on a plant genome: analysis of two *Arabidopsis transparent testa* mutations. *Plant Cell* 4:333–347
- Shu QY (2009) Turning plant mutation breeding into a new era: molecular mutation breeding. In: Induced plant mutations in the genomics era. FAO, Rome, pp 425–427
- Silkova OG, Dobrovolskaya OB, Dubovets NI et al (2006) Production of wheat-rye substitution lines and identification of chromosome composition of karyotypes using C-banding, GISH, and SSR markers. *Russ J Genet* 42:645–653
- Song XQ, Xia GM, Chen HM (2000) Chromosomal variation in long-term cultures of several related plants in *Triticinae*. *Acta Phytophysiol Sin* 26:33–38
- Stewart EA, McKusick KB, Aggarwal A et al (1997) An STS-based radiation hybrid map of the human genome. *Genome Res* 7:422–433
- Surzycki SA, Belknap WR (1999) Characterization of repetitive DNA elements in *Arabidopsis*. *J Mol Evol* 48:684–691
- Szakács É, Molnár-Láng M (2007) Development and molecular cytogenetic identification of new winter wheat-winter barley ('Martonvásári 9 kr1'-'Igri') disomic addition lines. *Genome* 50: 43–50
- Szakács É, Molnár-Láng M (2010) Identification of new winter wheat-winter barley addition lines (6HS and 7H) using fluorescence *in situ* hybridization and the stability of the whole 'Martonvásári 9 kr1'-'Igri' addition set. *Genome* 53:35–44
- Taski-Ajdukovic K, Nagl N, Miladinovic D (2010) Towards reducing genotype specificity in regeneration protocols after somatic hybridization between cultivated sunflower and wild *Helianthus* species. *Acta Biol Hung* 61:214–223
- Van Geyt JPC, Oleo M, Lange W, de Bock TSM (1988) Monosomic additions in beet (*Beta vulgaris*) carrying extra chromosomes of *Beta procumbens*. 1. Identification of the alien chromosomes with the help of isozyme markers. *Theor Appl Genet* 76:577–586
- Van Orsouw NJ, Hogers RC, Janssen A et al (2007) Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* 2:e1172
- Van Vactor D, Krantz DE, Reinke R, Zipursky SL (1988) Analysis of mutants in chaoptin, a photoreceptor cell-specific glycoprotein in *Drosophila*, reveals its role in cellular morphogenesis. *Cell* 52:281–290
- Vogel JP, Garvin DF, Mockler TC et al (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768
- Wade CM, Giulotto E, Sigurdsson S et al (2009) Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* 326:865–867
- Walter MA, Spillet D J, Thomas P et al (1994) A method for constructing radiation hybrid maps of whole genomes. *Nat Genet* 7:22–28
- Wang GX, Tang Y, Yan H et al (2011) Production and characterization of interspecific somatic hybrids between *Brassica oleracea* var. botrytis and *B. nigra* and their progenies for the selection of advanced pre-breeding materials. *Plant Cell Rep* 30:1811–1821

- Wardrop J, Snape J, Powell W, Machray G (2002) Constructing plant radiation hybrid panels. *Plant J* 31:223–228
- Wardrop J, Fuller J, Powell W, Machray GC (2004) Exploiting plant somatic radiation hybrids for physical mapping of expressed sequence tags. *Theor Appl Genet* 108:343–348
- Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562
- Weikard R, Goldammer T, Laurent P et al (2006) A gene-based high-resolution comparative radiation hybrid map as a framework for genome sequence assembly of a bovine chromosome 6 region associated with QTL for growth, body composition, and milk performance traits. *BMC Genomics* 7:53
- Weterings E, van Gent DC (2004) The mechanism of non-homologous end-joining: a synopsis of synopsis. *DNA Repair* 3:1425–1435
- Wi GS, Chung BY, Kim JS et al (2007) Effects of gamma irradiation on morphological changes and biological responses in plants. *Micron* 38:553–564
- Wu CH, Nomura K, Goldammer T et al (2008) A high-resolution comparative radiation hybrid map of ovine chromosomal regions that are homologous to human chromosome 6 (HSA6). *Animal Genetics* 39:459–467
- Xiang F, Xia G, Chen H (2003) Asymmetric somatic hybridization between wheat (*Triticum aestivum*) and *Avena sativa* L. *Sci China C Life Sci* 46:243–252
- Xue S, Zhang Z, Lin F et al (2008) A high-density intervarietal map of the wheat genome enriched with markers derived from expressed sequence tags. *Theor Appl Genet* 117:181–189
- Yamano S, Tsujimoto H, Endo TR, Nasuda S (2010) Radiation mutants for mapping genes and markers in pericentromeric region of chromosome 3B of Norin 26 wheat. *Wheat Inf Serv* 109:11–13
- Yang ZJ, Zhang T, Liu C et al (2008) Identification of wheat-*Dasypyrum breviaristatum* addition lines with stripe rust resistance using C-banding and genomic in situ hybridization. In: Appels R, Eastwood E, Lagudah E, Langridge P, Mackay M (eds) Proceedings of 11th international wheat genet symposium, Sydney University Press.
- Ya-Ping Y, Xiao C, Si-He X et al (2003) Identification of wheat-barley 2H alien substitution lines. *Acta Bot Sin* 45:1096–1102
- You Y, Bergstrom R, Lederman R et al (1997) Chromosomal deletion complexes in mice by radiation of embryonic stem cells. *Nat Genet* 15:285–288
- Zacharias M (1956) Mutationversuche an Kulturpflanzen. VI. Rontgenbes- trahlungen der So- jabohne (*Glycine soja* L.). *Sieb.et Zucc Zuchter* 11:321–338
- Zhang C, Yang Z, Gui X et al (2008) Somatic hybridization between *Brassica napus* and *Eruca sativa* mill. *Sheng Wu Gong Cheng Xue Bao* 24:793–802
- Zhang F, Wang P, Ji D et al (2011) Asymmetric somatic hybridization between *Bupleurum scorzonerifolium* Willd. and *Taxus chinensis* var. mairei. *Plant Cell Rep* 30:1857–1864
- Zhou C, Xia GM, Zhi DY, Chen Y (2006) Genetic characterization of asymmetric somatic hybrids between *Bupleurum scorzonerifolium* Willd. and *Triticum aestivum* L.: potential application to the study of the wheat genome. *Planta* 223:414–424

Chapter 13

FISHIS: A New Way in Chromosome Flow Sorting Makes Complex Genomes More Accessible

Sergio Lucretti, Debora Giorgi, Anna Farina and Valentina Grosso

Contents

13.1 High Sensitivity Tools to Investigate Cells and Genomes	320
13.1.1 Molecular Cytogenetics	320
13.1.2 Flow Cytometry and Flow Sorting: How Do They Work?	321
13.1.3 Next-generation Sequencing of Complex Crop Genomes	325
13.2 Challenging Complex Genomes	326
13.2.1 Flow Cytogenetics, the Chromosome Approach and Chromosome Genomics	327
13.3 Fishing New Chromosomes with FISHIS—Fluorescence <i>In Situ</i> Hybridization in Suspension	329
13.3.1 Labelling Floating Things	329
13.3.2 An Open Access to Potentially All Chromosomes in <i>Triticaceae</i>	337
13.3.3 The Bad and the Good: Limitations and Perspectives	340
References	344

Abstract Eukaryotic chromosomes can be studied either fixed on slides by use of cytogenetic techniques or through flow-cytometry. The latter enables not only the chromosome characterization but allows for the sorting and manipulation of individual chromosomes out of a complete genome, provided that target chromosomes have flow-cytometric distinctive features (e.g. size and/or DNA base content). These requirements represent a major constraint when the chromosomes of a given species are similarly sized, a common feature in plants. In wheat, this constraint can be overcome using special aneuploid mutants, each containing single arm pairs for a given chromosome of the complement, thus keeping a balanced genome composition while showing a complement with chromosome parts which are half-sized in respect to standard autosomes. However, such genotypes are available in hexaploid wheat (*Triticum aestivum*) exclusively in the background of Chinese Spring, a non-elite reference variety. Notably, aneuploids are not available for the most part of plants and animals. In order to overcome this major hurdle we have developed a reliable, fast and cost-effective method for Fluorescence labeling and *In situ* Hybridization

S. Lucretti (✉) · D. Giorgi · A. Farina · V. Grosso
ENEA—Italian National Agency for New Technologies, Energy and the Environment,
Casaccia Research Center, 00123 Rome, Italy
e-mail: sergio.lucretti@enea.it

of chromosomes In Suspension (FISHIS). The method makes use of fluorescent oligonucleotides of simple repetitive DNA sequences (SSR) or short DNA fragments, as probes, and allows for specific chromosome flow-sorting based on the hybridization pattern determined by the FISHIS probe. We have successfully applied the FISHIS methodology for flow karyotyping and sorting highly pure, single-type chromosome fractions of commercial varieties of bread and durum wheat and other *Triticeae* species. Moreover, the complete chromosomal sets corresponding to the two genomes (A and B) of durum wheat have also been clearly separated by the same FISHIS approach. Given the ubiquitous occurrence in eukaryotic genomes of the type of sequences used as probes, the abundance of their variants and their overall chromosome-specific distribution, their use in FISHIS experiments provides simple and fast access to individual chromosomes of virtually any eukaryotic genome, paving the way for gaining knowledge of great potential impact on basic and applied research aspects.

Keywords Plant flow cytometry · *In situ* fluorescent hybridization · Microsatellites · SSRs · Repetitive DNA sequences · Next generation sequencing · Wheat

13.1 High Sensitivity Tools to Investigate Cells and Genomes

13.1.1 Molecular Cytogenetics

In the “classic” cytological studies, chromosomes are fixed on a slide and act as the main subject of investigation in cytogenetics which mostly evaluates their number and size rearrangements. But many unseen changes happen at the genomic level and those can be investigated by the use of *in situ* hybridization techniques (ISH) where a nucleic acids molecular probe specifically hybridizes to the target DNA or RNA, thus pointing out researchers’ interest to molecular cytogenetics (Speel 1999). One of the most useful hybridization procedure makes use of fluorescent labeled nucleic acids probes, or Fluorescent *In situ* Hybridization (FISH), which has almost completely replaced ISH using radioisotope labeled DNA probes in chromosome and nucleus studies (Chester et al. 2010).

Molecular cytogenetic probes can be prepared from a number of specific sequences, from the smaller probes obtained from unique sequences, microsatellite and telomere DNA repetitive sequences, PCR amplification of locus specific DNA, to the larger probes like whole genomes (GISH), large genomic DNA sequences cloned into cosmids, bacterial artificial chromosomes (BACs), yeast artificial chromosomes (YACs), or generated by microdissection of chromosome arm specific bands, and BAC DNA libraries established by chromosome flow sorting. Single or multiple probes, each labeled with a different fluorochrome, can be used for *in situ* hybridization, resulting in multiple localizations and a more informative physical mapping of the sequences of interest (Jiming and Gill 2006).

13.1.1.1 *In Situ* Fluorescent Hybridization (FISH)

The “classic” FISH techniques operate on morphologically maintained chromosomes, cells or tissue sections, to detect specific nucleic acid sequences and gathering informations on their physical mapping and related gene activity at DNA, mRNA and protein level. FISH consistency has been recognized by a large set of experimental data and has made this technique a molecular diagnostic tool to trace or corroborate mutations and abnormalities at the gene or chromosome level. The common FISH procedure envisage the labeling of the nucleic acids probes (e.g. DNA or RNA sequences) by the incorporation of nucleotides which are coniugated with a fluorescent dye (direct labeling) or with a molecule which can be reconized by a fluorescent labeled antibody (undirect labeling with signal amplification possibilities). At first, the target sample (e.g. metaphase chromosomes or interphase nuclei) and a fluorescent probe are both denatured and mixed together into the hybridization buffer, then left to hybridize. The nucleic acids are allowed to re-anneal back into a double helix and the probe can then be seen as a fluorescent signal by microscopic observation; the sample DNA can be scored for the presence or absence of the signal, and spotted for its physical localization. FISH straightforward applications are found for the identification of typical marker chromosomes and chromosomal rearrangments, the scoring of microdeletions, the characterization of aneuploids. FISH specific labeling can be visualized not only on metaphase chromosomes, where DNA gets its maximum condensation, but on interphase nuclei as well allowing the study of genomic alterations avoiding the demand for cell cycle manipulation, metaphase induction and chromosome preparation. The usefulness of FISH and its suitability for analyzing gene localization, chromosome rearrangments and genome evolution has been witnessed by the impressive amount of methods, and results, branched from (Speel 1999; Kato et al. 2005; Heslop-Harrison and Schwarzacher 2011) (Fig. 13.1).

13.1.2 *Flow Cytometry and Flow Sorting: How Do They Work?*

A flow cytometer used to be a rather complex machine, but nowadays much efforts have been committed to make these instruments more user-friendly, and powerful at the same time. Nevertheless, a flow cytometer is still a multifaced machine made up with four combined systems: fluidics, optics, electronics and data analysis and management, and the latter is what makes a flow cytometer usable for researchers.

13.1.2.1 Flow Cytometry Analysis

A general operational description of a flow cytometer (Fig. 13.2) starts from the sample tube where particles in suspension are stored. The sample is injected through the sample inlet into the flow chamber, where chromosomes are aligned one-by-one by hydrodynamic focusing, that is, a liquid is injected into the flow chamber together

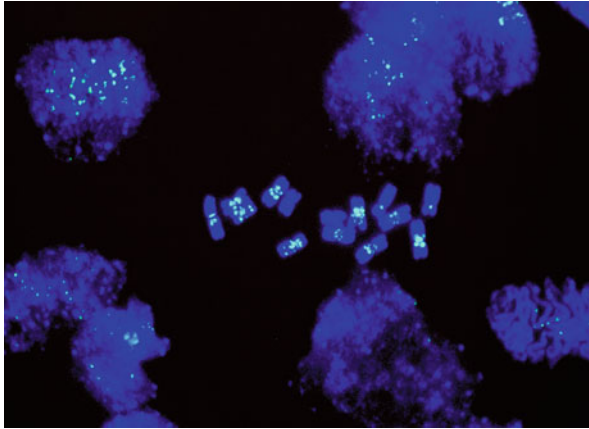


Fig. 13.1 The high classification power of the “classic” FISH on a slide: all the *Dasyvirum villosum* chromosomes but a couple, are hybridized with a specific pattern by the fluorescein conjugated oligonucleotide (GAA)₇-FITC allowing the discrimination of each chromosome type (Grosso et al. 2012). A peculiar banding pattern can also be observed on interphase nuclei which can be exploited for assessing the presence of a specific DNA sequence

with the sample, but at a much high pressure, forcing the sample core to reduce its diameter to approach the size of the chromosomes. Chromosomes are gently aligned and as they leave the flow chamber to air, they are hit by a strong excitation beam (e.g. a laser) which allows simultaneous fluorescent emissions and light diffraction from each single chromosome crossing the intersection point with the laser beam. This way, each chromosome gives information about its physical properties (forward and side light scatter: a rough estimation of dimension and texture, respectively) and biochemical features by specific staining with fluorochromes (e.g. DNA content with DAPI: 4-4', 6 diamidino-2-phenylindole). Light scatters and fluorescence signals are captured by specific sensors which translate their intensity to electric pulses and digital signals to be processed by the data analysis and decision software. This way, all particles emitting a signal are analyzed and cataloged by the system computer software to discriminate subpopulations within the sample (Fig. 13.3).

13.1.2.2 Flow-sorting

The flow cytometer is the only instrument which can couple a high analytical capability with a high-end preparative option, namely flow-sorting (Fig. 13.2). This unique feature is largely employed in biomedical sciences and diagnostics, but its use is becoming more and more diffuse in plant biotechnology and genomics as well (Galbraith 2010). Flow sorting enables the isolation of a highly pure subpopulation of particles (chromosomes) at a purity which can reach up to 99%, but usually > 95% is a common feature. The particles can be collected in an axenic way and are then made available for morphological or genetic examination, as well as functional assays and/or therapeutics.

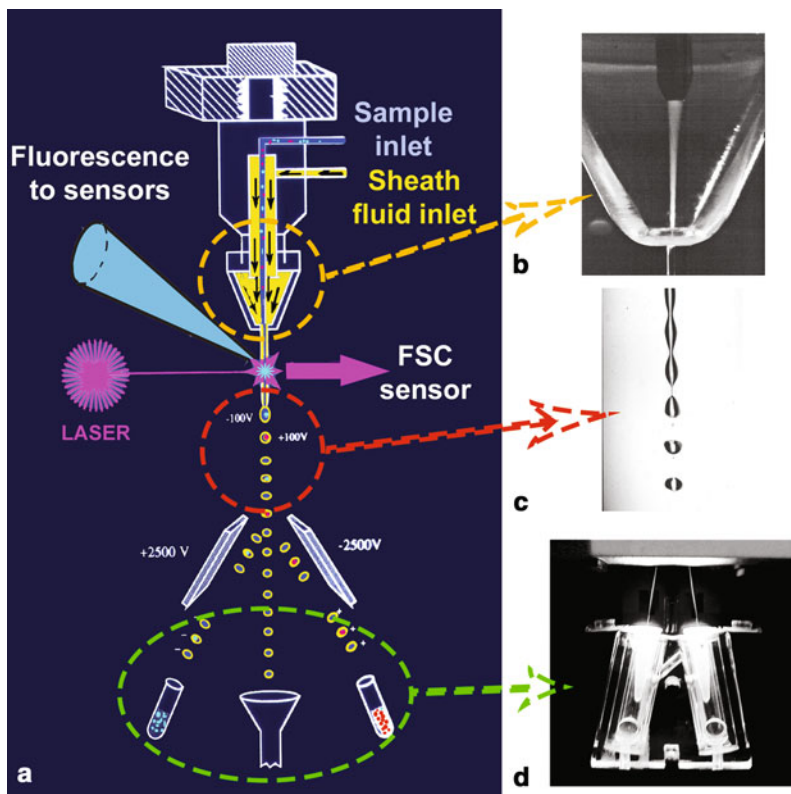


Fig. 13.2 a) Schematic drawing of a flow cytometer and cell sorter-in-air. The sample (*light blue color*) is injected at a constant rate into a flow chamber where a sheath fluid (*yellow color*) at higher pressure gently forces the particles into a single line (hydrodynamic focusing). In **b**) an enlarged image of the flow-chamber shows the sample internal core which escapes from the nozzle tip and intercepts the laser beam (interrogation point) in air. At the interrogation point, scatter signals and fluorescent emissions are generated and all of them are selectively collected by the digital sensors; per each single particle a graphic representation of the signals is presented on the system computer screen and the operator can decide the range and combination of values related to the particle of interest to be sorted out of the whole sample. Then, an electric pulse is sent to the liquid stream in coincidence with the formation of the droplet containing the particle of interest, as shown in **c**). The particle in-air retains its charge and is deviated from the mainstream once it enters the stationary electromagnetic field generated by two electrodes along its way. In **d**) chromosome-containing droplets are sorted at high rate, so resembling a continuum (sort rate 1,000 drops/sec). All the flow sorted particles can be collected in specific receptacles (e.g. low DNA binding Eppendorf tubes) in axenic conditions and ready for subsequent molecular and/or biological manipulations

The sorted sample is collected into a liquid portion of the sheath fluid (SF), as schematically presented in Fig. 13.2. According to the available sorting method, the SF amount can span from few nanoliters to tens of milliliters which greatly affects the end use of the sorted fractions, since collecting particles into a large SF volume would add a concentration step to sample processing (Galbraith and Lucretti 2000). For molecular applications, the smaller the volume, the easier it is to use the flow-sorted samples. The most widespread method of sorting chromosomes is by electrostatic

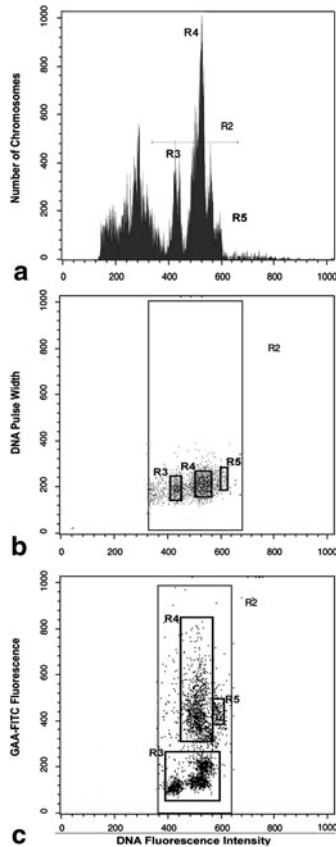


Fig. 13.3 Flow cytometry data are saved on a cell-by-cell basis (list-mode) generating a matrix of values according the particle features (parameters) measured and stored in a standard format (FCS) which allows for an easy access to analyze and compare experiments. Usually, each single parameter can be shown and analyzed on a computer screen as a single distribution of fluorescence values related to the content of the stained particle components (histogram) or as a pair of parameters (dot plot), and each dot on the display shows an analytical event (particle of interest). These graphic presentations of the analytical data let us doing an easy setting of the regions of interest to be drawn around the clusters of the candidate particles of interest for flow-sorting (sorting gates). In **a**, a DAPI (4,6-diamidino-2-phenylindole) stained chromosomes flow karyotype of *T. durum* Creso is presented showing the distribution frequency of the fluorescence emitted from single chromosomes analyzed one-by-one. This histogram allows the distinction of three regions only (*R2*, all chromosomes; *R3*: 1A, 6A; *R4*: 3A, 5A, 4A, 2A, 7A, 1B, 2B, 4B, 5B, 6B, 7B; *R5*: 3B). In **b**, a dual parameter analysis (dot plot) is presented which combines the evaluations from the DNA histogram plus the a second parameter related to size (DNA Pulse Width) measuring the duration of the pulse of the DNA fluorescence emission per each particle. Even so, the chromosomes detection does not improve and three regions are still shown. In **c** a new level of characterization has been achieved by FISHIS labeling of chromosomes with a molecular fluorescinated probe (GAA)⁷, and most of the Creso chromosomes can be located (*R2*: all chromosomes; *R3*: genome A all chromosomes and internal sub-regions; *R4*: genome B all chromosomes and internal sub-regions; *R5*: chromosome 3B. See Fig. 13.8 for a full characterization of all the sub-regions)

deflection of charged droplets in air as shown in Fig. 13.2. A sorting gate is drawn on the dot plot (e.g. regions R3-R4-R5) visualized in real-time by the computer (Fig. 13.3) which is translated by the machine in charging sorting-pulses. These pulses are synchronized with the droplet formation by a piezoelectric transducer which vibrates the flow chamber and generate a controlled breaking down of the fluid in air. All the analytical and sorting procedures are run in real-time in microseconds in respect to the milliseconds the particle needs to cover the entire distance from the laser interception point to the droplets breaking point, at which the charging pulse is sent by the instrument. For a nozzle tip of 70 μm in diameter, a typical droplet size is of 1–2 nl, depending on SF nature and speed, and droplet formation frequency (Galbraith and Lucretti 2000). The sorting step is described in Fig. 13.2, and the chromosomes of interest can be collected into any device which is needed; the unwanted sample stays on its way to the waste tank.

13.1.3 Next-generation Sequencing of Complex Crop Genomes

After the Sanger sequencing (Sanger et al. 1977) has achieved a number of significant feats, including the whole human genome sequencing (Lander et al. 2001), some faster and more and more affordable technologies came to the scenario, named as second- and third-generation sequencing technologies or “Next Generation DNA Sequencing—NGS” (Cook and Varshney 2010; Metzker 2010). NGS promised and allowed new opportunities to speed up genome sequencing and to investigate the primary structure of large and complex genomes (Edwards et al. 2013). As of now, a list of 35 higher plants have been sequenced (http://en.wikipedia.org/wiki/List_of_sequenced_plant_genomes), including both methods.

NGS technologies have burst the development of new genomic resources for many and recalcitrant plant species, but till now only *Arabidopsis* (157 Mbp) and *Oryza* (466 Mbp) have a fully available and published sequencing (Kaul et al. 2000; Chen et al. 2013). Crop genomes are often of large size, and more often they are polyploids (e.g. wheat, corn, brassica, soja, potato) containing a large amount of repetitive sequences which proved to be recalcitrant to assembly methods, and this holds true even for the “gold standards” *Arabidopsis* and *Oryza* (Claros et al. 2012).

All the more, this applies for the complex genomes of wheat (*T. durum* and *T. aestivum*), where the huge size of the genome (about 12 and 17 Gbp, respectively) and the elevated complexity coupled with a particularly high prevalence of gene duplication due to their more recent polyploidy events, represent difficult obstacles that have hindered the development of reliable reference sequences and genome-wide SNP resources. An international initiative is actively pursuing the aim to generate a reference genome sequences for wheat, but the sheer size of the genome necessitates alternative approaches to whole-genome *de novo* sequencing (<http://www.wheatgenome.info/wiki/>).

13.2 Challenging Complex Genomes

Higher plants have successfully colonized the planet with a rapid diversification which corresponds not only to an extremely changeable phenotype but reflects a highly variable DNA content, spanning from the 64 Mbp genome of *Genlisea* (corkscrew plants) to the 149 Gbp genome of *Paris japonica* (Pellicer et al. 2010) with an increase of more than 2,300 fold. Some changes happen at the individual level between plants belonging to the same species with the same chromosome number but showing up to 40 % difference in DNA content (Greilhuber 2005; Šmarda et al. 2008).

A number of studies have shown that the entire genetic features of a species are not at all encompassed into a single genome. Accordingly, a new concept has been raised to define more complex and evolutive entities for each species, namely the pangenome and core genome (Morgante et al. 2007; Hansey et al. 2012). This complexity of the angiosperm genome relates to many phenomena that have shaped plant evolution, speciation and diversity within each species, (Paterson et al. 2000; Bento et al. 2011; Wang et al. 2012). Gymnosperms do not make an exception to this, as it is in the case of Conifers where the more diffuse and relevant genera have a huge haploid genome size spanning from 18 to 35 Gbp, featuring a different gene organization from angiosperm and a large amount of retrotransposons (Mackay et al. 2012). A large appraisal of the diverse genetic mechanisms contributing to plant genome size variations is mandatory to allow assembling the huge amount of sequencing data which can be provided by NGS. The most relevant contribution to plant genome size is due to repetitive sequences, most of them related to transposable elements which presence extent from 5.6 % in the rice small genome to 70 % in barley and 85 % in corn and wheat large genomes (Schnable et al. 2009). Their abundance and short repetitive sequences texture are one of the main obstacle to assemble sequencing data (Henson et al. 2012). A further issue in assembling is the heterozygosity of most plants which complicates the clustering when physical markers are used to drive NGS. A solution to this problem is to sequence homozygous derivatives as in grape (Velasco et al. 2007) or potato (Consortium TPGS 2011). A special feature of plants is the large number of polyploidy species where two or more genomes provide more evolutionary plasticity (Comai 2005) while confounding NGS assemblers because of the difficulty to distinguish among homeologous chromosomes sequences, gene doubling and tandem repeats (Treangen and Salzberg 2012; Henson et al. 2012; Bradbury et al. 2013). A possible solution to circumvent such a situation is the use of dihaploids (as in the case of the tetraploid potato; Consortium TPGS 2011) or to use the chromosome approach, as in bread wheat (Feuillet et al. 2011).

More general repetitive sequences are present both in plants and other organisms which are responsible for generating low complexity regions, still difficult to interpret and assign during assembling, especially in *de novo* sequencing (Page et al. 2013). Copy-number variation (CNV) in plants are likely to be an important factor in genome size variations and genetic functionality as already substantiated in human studies (Choy et al. 2010) and described in barley for an increased boron tolerance (Sutton et al. 2007), in rice for tolerance to hypoxic conditions (Xu et al. 2006) and in *Arabidopsis* for improved adaptative capacity (DeBolt 2010). Further

examples of CNV are the rRNA genes packed in rDNA units, accounting for up to 10 % of the whole genome (Pruitt and Meyerowitz 1986). Lots of repetitive identical small sequences are present in satellites, which are mostly located at the centromeres and constitutive chromatin, and microsatellites (or SSRs: simple sequence repeats) where short tandem repeats (e.g. 1-8 nucleotides) are gathered in the kbp range, mainly at subtelomeric regions, but in a non-random fashion allowing chromosome discrimination (Kalia et al. 2011). Telomeric sequences are also made of short repeats, all very similar to the sequence TTTAGGG and present at kbp amounts at the physical end of each chromosome arm (Mizuno 2008), and both the lack of genetic markers and their repetitive structure complicate their assembly.

13.2.1 *Flow Cytogenetics, the Chromosome Approach and Chromosome Genomics*

Flow cytogenetics (FCY) is based on the use of chromosomes in suspension which can be analyzed (flow karyotyping) and sorted with a flow cytometer. FCY can evaluate large numbers of chromosomes producing flow karyotyping of high statistical value, and allows the development of the “chromosome approach” beneficial to deconvolute complex genomes into single chromosomes containing only a well-defined fraction of the whole genome and therefore easier to deal with (Doležel et al. 2009).

Chromosomes can be isolated only at the metaphase stage of the cell cycle, prefiguring the need for a very efficient procedure for cell cycle synchronization and chromosome isolation from easy to handle meristematic tissues. In plants, several attempts have been carried out to develop such an ideal system by using different kinds of explants, such as protoplasts, cell cultures, hairy roots, and fast-growing roots. In one of the earliest reports on plant flow cytogenetics, Delaet and Blaas (1984) described the first chromosome flow sorting in *Haplopappus gracilis* ($2n = 4$) from crude protoplasts isolated from *in vitro* synchronized cell suspensions. Fresh synchronized protoplasts were utilized as the main experimental material until the beginning of the '90s (Conia et al. 1987; Arumuganathan et al. 1991; Wang et al. 1992) when a new procedure was developed using *V. faba* seedlings with fast growing roots (Doležel et al. 1992). A high mitotic index was achieved by a double-step cell cycle synchronization and a large number of intact chromosomes were released in suspension after root fixation with formaldehyde and fine chopping the root tips into a lysis buffer. This isolation method has several advantages such as the simple handling of root seedlings, the stability of karyotypes, and conformity of chromosomes, and the easy manipulation of the cell cycle to induce synchronicity in DNA synthesis and metaphase arrest. Formaldehyde-fixed chromosomes have proved to be suitable for flow sorting (Lucretti et al. 1993), DNA amplification (Macas et al. 1993) and molecular manipulation (Macas et al. 1995) setting the basis for the chromosome approach (Lucretti and Doležel 1995) (Fig. 13.4).

Methods for chromosome isolation and flow sorting have kept developing, and nowadays they are reported for 22 plant species including major cereal crops and wild

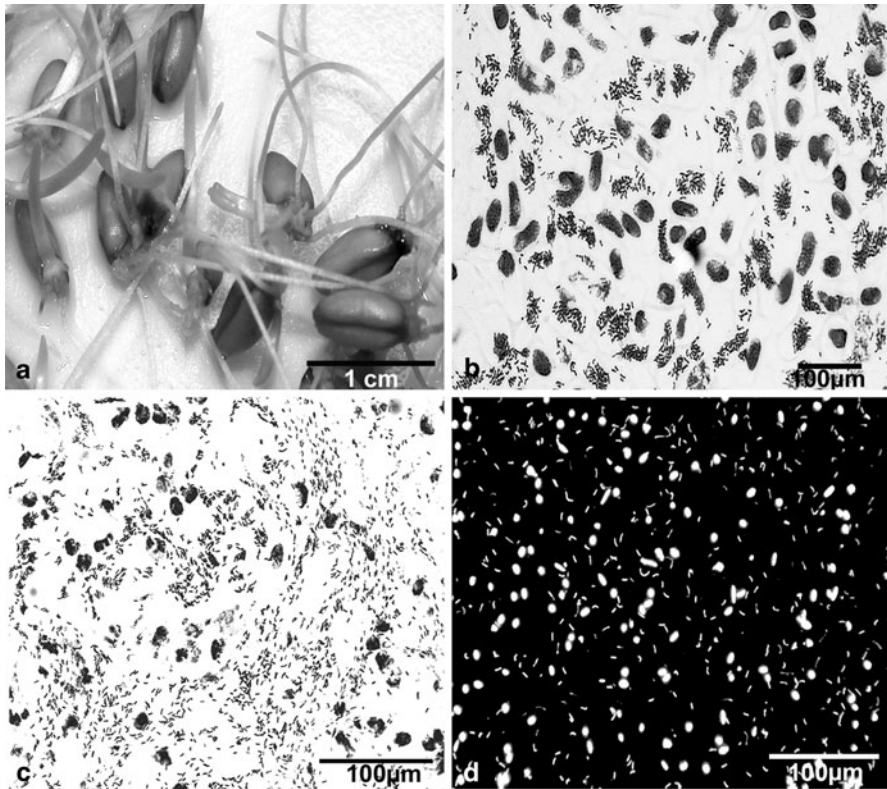


Fig. 13.4 From growing roots to chromosomes in suspension. Roots from seed-propagated plants are the explants of choice for chromosome isolation (a). Seeds are inhibited, germinated and grown for few days onto a liquid medium with air bubbling. b and c, root tip fast growing meristematic cells are synchronized by hydroxyurea at the beginning of the DNA synthesis, then they are released from the block by removing the drug and then stopped at metaphase by adding a mitotic spindle inhibitor. This dual-step cell cycle manipulation allows for a mitotic index of 50–60 %, which is mandatory for obtaining a high quality chromosome suspension (b = *T.aestivum* and c = tetraploid *Triticale* Oceano metaphases, respectively). Treated roots are fixed in a formaldehyde buffer which makes the root tissues hard but fragile, and suitable for homogenization by sharp blades. Intact chromosomes are released into the isolation buffer, ready for further treatments and/or flow karyotyping and sorting. In d, a suspension of wheat chromosomes and nuclei stained with DAPI is shown

relatives (Doležel et al. 2012; Grosso et al. 2012). Chromosome genomics, that is genome analysis using chromosome-based approaches and flow cytogenetics, is now growing in significance and applicability as new methods in cellular and molecular biology and genomics are developing (Doležel et al. 2007, 2012).

The International Research Initiative for Wheat Genome Sequencing Consortium IWGSC (<http://www.wheatgenome.org>) has chosen the chromosome approach to tackle the complexity of the 17 Gbp of the bread wheat polyploid genome, made by three homeologous genomes containing over 80 % repetitive DNA sequences (Smith and Flavell 1975). The Consortium approved the use of aneuploid mutants

(*T. aestivum* Chinese Spring double ditelosomic lines; Sears and Sears 1978) to isolate and sequence single chromosome arms that differ in size from the standard complement (Gill et al. 1999). Large insert bacterial artificial chromosome (BAC) libraries (Šafář et al. 2004) are being prepared from Chinese Spring sorted arms to sequence and construct a high quality assembled bread wheat genome. These aneuploid lines are not easily available, even for other wheat varieties. Under these circumstances, the chromosome approach could not have been extended to other agronomically important crops.

13.3 Fishing New Chromosomes with FISHIS—Fluorescence *In Situ* Hybridization in Suspension

13.3.1 Labelling Floating Things

Nowadays, the chromosome approach has reached a high level of sensitivity and integration with chromosome genomics (Doležel et al. 2012). But plant chromosome flow sorting is still based on fluorescence signals to discriminate chromosomes by DNA content and chromosome size. An innovative strategy for precisely labeling individual chromosomes in suspension would be a desirable achievement since it would improve the resolution power of a flow cytometer allowing a new way of characterizing, and therefore flow sorting, an all new set of chromosomes. Thus, the chromosome approach would offer to genomics a wider gene pool and many more species to work with (Doležel et al. 2012). Since the analytical capability of a flow cytometer and cell sorter uses fluorescence signals to characterize chromosomes, we focus our attention on a high discriminatory FISH labeling strategy involving *in situ* fluorescent hybridization exploiting either chromosome-specific DNA sequences, or repetitive DNAs with chromosome-specific distribution patterns. The way in which a standard flow cytometer looks at the sample, as a speedy fluorescent spot of light, rules out the use of low fluorescence intensity single-copy sequences as hybridization probes, so far. Yet satellites, microsatellites (or SSRs) and transposable elements are distributed throughout genomes (Sharma and Raina 2005) in a non-random fashion and are widely used in cytogenetic characterization and identification of individual chromosomes (Pedersen and Langridge 1997; Cuadrado and Jouve 2010; Payseur et al. 2011).

The standard FISH protocol relies on the denaturation of the probe and target DNAs and their annealing under sufficiently stringent conditions to warranty specific and reproducible hybridization (Fig. 13.5a).

Much effort was devoted to apply the classic FISH labeling procedure to nuclei and chromosomes in suspension, first in human (Trask et al. 1985), and then in plants (Pich et al. 1995; Ma et al. 2005) but results were largely unsatisfactory. Recently Brind'Amour and Lansdorp (2011) have succeeded to label human chromosomes in suspension with synthetic nucleotides, but flow sorting was not reported. As a matter

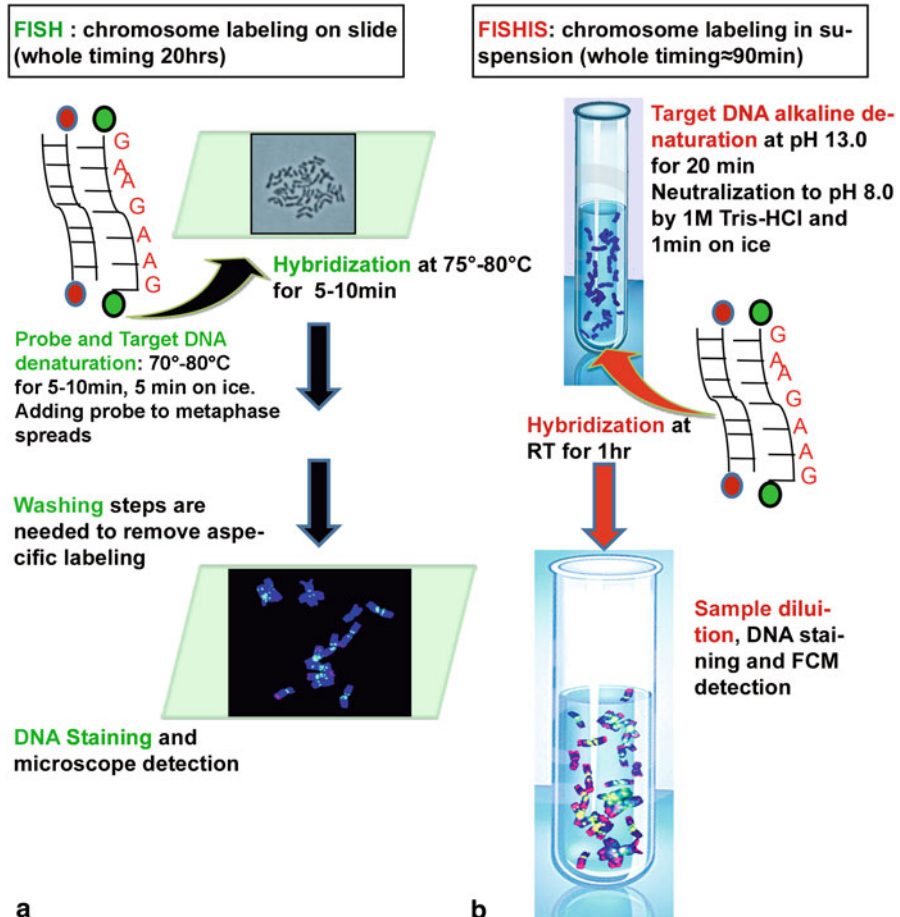


Fig. 13.5 Comparing FISH and FISHIS: **a** the FISH procedure is simply described showing the main steps which generate a specific hybridization signal on metaphase-spread chromosomes, underlining the harsh hot denaturing and washing steps, both of which affect sample firmness and accessibility. **b** FISHIS acts on particles in suspension (chromosomes and nuclei) in a very short time in comparison to FISH, to deliver a precise hybridization pattern without affecting sample integrity and accessibility, which is very suitable to flow cytometry analysis and sorting

of fact, chromosomes in suspension are inclined to dissolve or clump together when exposed to the harsh heat denaturation settings required for FISH. From some of our previous experiments (Lucretti et al. 1999) we concluded that a gentle procedure for FISH labeling could be possible using an alkaline denaturation step while avoiding any washing during the sample setting up. We have called this method FISHIS, or fluorescent *in situ* hybridization in suspension (Fig. 13.5b).

A New FISH Labeling Technique for Nuclei and Chromosomes in Suspension
FISHIS combines the high-throughput and preparative capabilities of flow cytometry

and the specificity of FISH labeling required to sort pure samples of chromosomes and nuclei. FISHIS is an easy method that removes any centrifugation and hot denaturation standard FISH steps, mixing all reagents together after an alkaline DNA denaturation (Ageno et al. 1969). Synthesized fluorescently labeled DNA repetitive sequences (e.g. SSRs) are used as one of the standard probes, further simplifying the process for FISHIS labeling (Raap et al. 1986). Our aim was to develop an effective procedure able to discriminate, purify and flow sort: (i) all the chromosome belonging to a single genome from the homeologous ones in polyploid pasta and bread wheat; (ii) several single-type chromosomes of bread and pasta wheat, and (iii) all the single-type chromosomes of the entire chromosome complement of the diploid wild wheat relative *Dasypyrum villosum* (L.).

Setting Up Suspensions of Plant Chromosomes and Nuclei We describe a procedure which gives good results in wheat and *D. villosum* and could easily be fitted to isolate chromosomes by other plant species with only minor modifications (Table 13.1).

Grains are soaked in aerated water for 8–24 h and germinated on moist filter paper for 2 days in the dark at $19 \pm 1^\circ\text{C}$ (root length 2–3 cm). Cell cycle synchronization of the root tip meristematic cells was achieved by exposure to 1.25 mM hydroxyurea for 18 h, followed by immersion in aerated Hoagland's solution (Doležel et al. 1999), for 4 h for pasta wheat and *D. villosum* and for 4.5 h for bread wheat. Cell division was blocked at metaphase by a 2 h treatment with 2.5 μM amiprofos-methyl (this step should be adapted to the species), and the resulting metaphase-arrested chromosomes were elongated and dispersed within the cytoplasm by an overnight incubation in ice water (standard procedure in grains, also recommended for other plant species). To prepare chromosome suspensions (Doležel et al. 1999), roots were excised and fixed in 3% (v/v) formaldehyde in 1x Tris-HCl pH7.5 at $5 \pm 1^\circ\text{C}$ for 20 min and rinsed three times in 1x Tris-HCl pH 7.5 at $5 \pm 1^\circ\text{C}$ for 5 min. The distal 1 mm of a set of 100 roots was homogenized in 1 ml LB01 lysis buffer (Ultraturrax T10 with G5 generator, IKA, Germany) at 13,500 rpm per 12 s into polystyrene tubes (this step should be adapted to the species) and the resulting homogenate filtered through a 36 μm nylon mesh to remove debris (Gualberti et al. 1996). Nuclear suspensions were obtained by the same procedure omitting all cell cycle synchronization steps.

Fluorescent DNA Probes A list of the repetitive sequences used by FISHIS labeling is given in Table 13.2. Ready-synthesized probes were end-labeled by FITC or Cy3 and HPLC desalted and resuspended at 1 $\mu\text{g}/\mu\text{l}$ in 10 mM Tris, 1 mM EDTA, according to manufacturer's instructions.

The 18S-5.8S-26S rDNA clone pTa71 (Gerlach and Bedbrook 1979) was labeled with FITC or Cy3 by nick-translation using standard kits (Nick Translation Mix, Roche) following manufacturer's instructions. The optimal size for the probe was experimentally found to be 200–500 bp (Fig. 13.6).

FISHIS Labeling

i) Denaturation of DNA by alkali

A total of 4 μl 4N NaOH were added to 150 μl of suspended chromosomes (2×10^6 chromosomes/ml LB01). The pH of the solution was measured with an

Table 13.1 Relevant parameters to optimize for the isolation of *Triticeae* chromosomes in suspension

Procedure (hours)	Species (genome)									
	<i>T. monococcum</i> (AA)	<i>T. urartu</i> (BB)	<i>T. durum</i> Creso (AA+BB)	<i>T. durum</i> Cappelli (AA+BB)	<i>T. aestivum</i> Provinciale (AA+BB+DD)	<i>T. aestivum</i> CSdDT5A (AA+BB+DD)	<i>S. cereale</i> Nikita (RR)	Triticale Oceania (AA+BB+RR)	<i>H. vulgare</i> Cometa (HH)	<i>D. villosum</i> D200 (VV)
Imbibition	8	8	8	8	8	8	8	8	8	24
Germination	48	48	48	48	48	48	48	48	48	48
Synchronization (mM HU)	18 (1.25)	18 (1.25)	18 (1.25)	18 (1.25)	18 (1.25)	18 (1.25)	18 (1.25)	18 (1.25)	18 (1.25)	18 (2)
Recovery	4	4	4.5	4.5	4	4	4	4	4	4
Blocking (μ M APM)	3 (2.5)	2.5 (2.5)	2 (2.5)	2.5 (2.5)	2 (2.5)	2 (2.5)	3 (2.5)	3 (2.5)	3 (2.5)	2.5 (2.5)
Chromosome spreading at 5°C	18	18	18	18	18	18	18	18	18	18
Reference	Lee et al. 2004 This paper	This paper	Kubaláková et al. 2005	Kubaláková et al. 2005	Vrána et al. 2000	Vrána et al. 2000	Vrána et al. 2000	Lee et al. 2004 This paper	Lysák et al. 1999	Grosso et al. 2012

Table 13.2 List the probes which can be used for FISHIS labeling

Probe	Species (genome)									
	<i>T. monococ- cum</i> (AA)	<i>T. urartu</i> (BB)	<i>T. durum</i> Creso (AA + BB)	<i>T. durum</i> Cappelli (AA + BB)	<i>T. aestivum</i> Provinciale (AA + BB + DD)	<i>T. aestivum</i> CSdDT5A (AA + BB + DD)	<i>S. cereale</i> Nikita (RR)	<i>Triticale</i> Oceania (AA + BB + RR)	<i>H. vulgare</i> Cometa (HH)	<i>D. villosum</i> D200 (VV)
(AAC) ₅	ND- FISH + (3) FISHIS + (3)	ND- FISH + (4) FISHIS + (4)	ND- FISH + (AA = 3)(BB = 7) FISHIS + (AA = 3)(BB = 7)	NA	NA	NA	ND- FISH + (?) FISHIS + (?)	ND- FISH + (?) FISHIS + (?)	FISHIS + (?) NA	ND- FISH + (5) FISHIS + (5)
(ACT) ₅	ND- FISH + (1)	NA	NA	NA	NA	NA	NA	NA	NA	NA
(AGG) ₅	-	FISHIS + (?) FISHIS + (1)	FISHIS + (?) FISHIS + (1)	FISHIS + (?) FISHIS + (1)	FISHIS + (?) FISHIS + (1)	FISHIS + (?) FISHIS + (1)	NA	NA	NA	FISHIS + (?)
(AAT) ₇	ND- FISH + (1)	FISHIS + (?) FISHIS + (1)	ND-FISH + (?) FISHIS + (?)	NA	NA	ND-FISH + (2)	NA	NA	NA	ND- FISH + (?) FISHIS + (?)
(AT) ₁₂	ND- FISH + (1)	NA	NA	NA	NA	NA	ND-FISH + (3) FISH + (?)	NA	NA	NA
(ATT) ₈	ND- FISH + (1)	NA	FISHIS + (?)	NA	NA	NA	ND-FISH + (?) FISH + (?)	NA	NA	NA
(CA) ₁₀	NA	NA	ND-FISH + (?) FISHIS + (1)	NA	NA	NA	NA	NA	NA	NA
(GAA) ₇	ND- FISH + (1)	ND- FISH + (2) FISHIS + (1)	FISHIS + (AA = 6; BB = 7)	FISHIS + (AA = 6; BB = 7)	FISHIS + (AA = 6; BB = 7)	FISHIS + (AA = 6; BB = 7)	FISHIS + (3) FISH + (3)	ND- FISH + (?) FISH + (?)	NA	FISHIS + (6) ND- FISH + (6)
(GACA) ₄	NA	NA	ND-FISH + (3) FISHIS + (3)	NA	NA	NA	ND- FISH + (?) FISH + (?)	NA	NA	ND- FISH +

Table 13.2 (continued)

Probe	Species (genome)									
	<i>T. monococcum</i> (AA)	<i>T. urartu</i> (BB)	<i>T. durum</i> Cresco (AA + BB)	<i>T. durum</i> Cappelli (AA + BB)	<i>T. aestivum</i> Provinciale (AA + BB + DD)	<i>T. aestivum</i> CSdDT5A (AA + BB + DD)	<i>S. cereale</i> Nikita (RR)	<i>Triticale</i> Oceania (AA + BB + RR)	<i>H. vulgare</i> Cometa (HH)	<i>D. villosum</i> D200 (VV)
(CCG) ₇	FISH + (1)	NA	NA	FISH + (?)	NA	NA	NA	NA	NA	NA
(AG) ₁₂	NA	NA	ND- FISH + (BB = 4)	NA	FISH + (?)	FISH + (?)	ND-FISH + (?) FISH + (?)	NA	NA	ND- FISH + (1)
(CAT) ₅	NA	NA	FISH + (BB = 4)	NA	NA	NA	FISH + (?)	FISH +	NA	FISH + (1)
(TTAGGG) ₅	NA	NA	FISH + (BB = 1)	NA	NA	NA	NA	(?)	NA	FISH + (1)
pTa71 (Gerlach and Bedbrook 1979)	NA	NA	FISH + (BB = 1)	NA	NA	NA	NA	NA	NA	ND- FISH + (1)
pSc 119.2 (Bedbrook et al. 1980)	NA	NA	FISH + (14)	NA	NA	NA	NA	NA	NA	FISH + (1)
pHv62 (Li et al. 1995)	NA	NA	FISH + (2)	FISH + (2)	NA	NA	NA	NA	NA	FISH + (1)
Reference	Molnár et al. 2011	This paper	Kubaláková et al. 2005	Giorgi et al. 2013	Vrána et al. 2000;	Vrána et al. 2000;	Vrána et al. 2003	Cuadrado et al. 1998	Cuadrado et al. 2007	Gosso et al. 2012
	This paper		2013	2013	2013	2013	et al. 1998	et al. 1998	Cuadrado et al. 2008	Giorgi et al. 2013
			2013	2013	2013	2013	et al. 2002	et al. 2002	Cuadrado et al. 2010	et al. 2013
			2013	2013	2013	2013	et al. 2008	et al. 2008	Carmona et al. 2013	

Number of labeled chromosomes in brackets NA Non Assessed

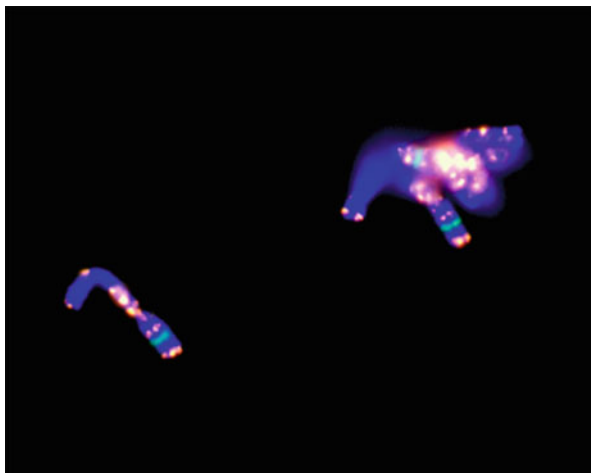


Fig. 13.6 *T. durum* Creso chromosomes were dual labeled by FISHIS with the oligo (GAA)₇-Cy3 (red signal) and the DNA probe pTa71 which was incorporated with dTTP-FITC nucleotides by nick-translation (green signal). Dual labeling FISHIS showed a well defined pTa71 band at the secondary constriction, thus indicating one of the two satellite wheat chromosomes, 1B or 6B. The joint presence of the GAA composite banding pattern allowed the identification of the two chromosomes in focus as 1B both (see Fig. 13.7)

ISFET micro pH probe inserted into the sample tube. The denaturation treatment was set at pH 13 for 20 min, followed by the addition of 1 M Tris-HCl pH 7.4 and maintaining the suspension on ice for 1 min to return to pH 8.0.

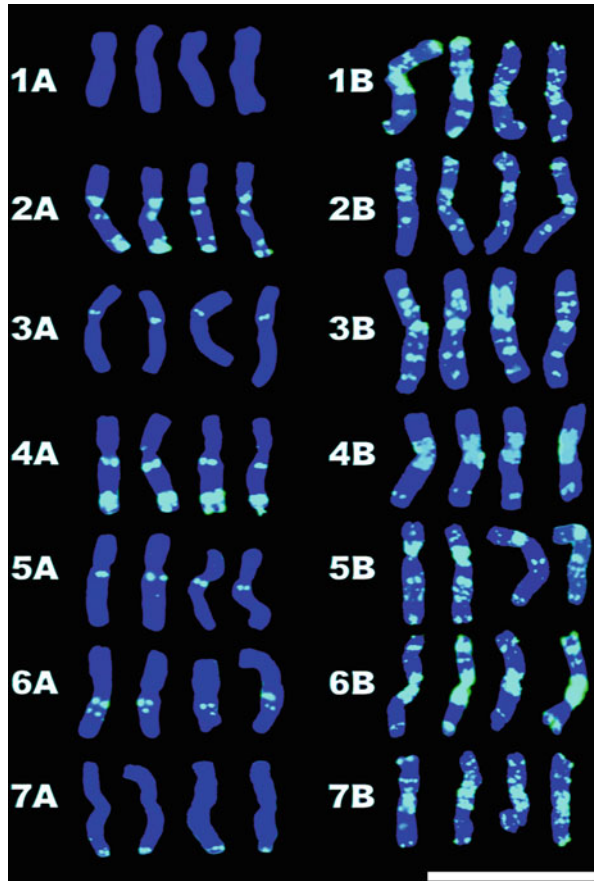
ii) Fluorescent labeling

The oligonucleotides were dissolved in 2XSSC at the concentration of 1.5–3.0 ng/ μ l and added to the sample at 160 ng/ml; we found this concentration adequate for labeling all the samples we tested. The labeling reaction was for 1 h at room temperature without any washing or centrifugation. After hybridization, samples were diluted 1:1 with LB01, counter-stained with 7 μ M DAPI and analyzed by flow cytometry. Each standard sample was 300 μ l (final volume), but the total sample volume can easily be scaled up to 1 ml, or more. Chromosome identification could be carried over directly by fluorescence microscopy, using 4 μ l chromosome suspension mounted in 30 % LB01 and 70 % Vectashield (v/v) (Vector Labs, Burlingame, CA) containing 7 μ M DAPI. In Fig. 13.6, a dual-labeling FISHIS with (GAA)₇-Cy3 and pTa71-FITC on *T. durum* Creso chromosomes in suspension is showed, with satellite regions easily reconizable at the secondary constriction.

Flow Cytometry and Chromosome Sorting FISHIS labeling proved to be of high intensity (Fig. 13.7), and consistent with the patterns obtained after standard FISH labeling on slide (Pedersen and Langridge 1997; Kubaláková et al. 2005).

Our “well trained” dual laser FACS Vantage SE flow cytometer (BD Bioscience, San Jose, CA) jet-in-air cell sorter allowed us to discriminate all the fluorescence

Fig. 13.7 The FISHIS (GAA)₇ labelling pattern in *T. durum* Cresco sorted chromosomes. Wheat chromosomes were FISHIS labeled with (GAA)₇-FITC and four chromosomes were collected for each autosome. The FISHIS GAA banding pattern correlated well with the previous patterns presented in Petersen and Langridge (1997) and in Kubaláková et al. (2005) and all the chromosomes were identified. The two homeologous genomes A and B were easily discriminated by their banding pattern and its intensity, being the chromosomes from the A-genome less labeled in respect to the B ones. Chromosomes DNA was counterstained with DAPI (DNA labelling, *blue color*). Bar = 10 μm. (From doi: doi:10.1371/journal.pone.0057994.s002)



signals from FISHIS samples (for details: see Giorgi et al. 2013). The chromosome suspension was easily run through a 70 μm nozzle flow tip at 23 PSI sheath pressure. Sample rate was set to 200 ÷ 1,000 particles/s by the step motor-driven 1 ml syringe. Usual sorting rate was 5–20 chromosomes/s in a dual-sorting mode. Sorted chromosomes were collected either on glass slides for real-time purity checking, or into low binding DNA tubes (Eppendorf LoBind, Germany) for further processing. The primary analysis gate was set on a dual parameter dot plot comprising particle size (Forward Scatter) versus DNA content (DAPI fluorescence) to role out debris and chromosome aggregates from intact chromosomes. The resulting dot plot of DNA integral fluorescence emission (FL1A) versus (GAA)₇ fluorescence labeling showed a much higher number of sub-populations (Fig. 13.3). In Fig. 13.8, sorting windows were drawn on the fluorescence dot plot of DNA content (FL1A) versus (GAA)₇ fluorescence labeling (FL3H); nucleotides were labeled by FITC (green fluorescence) or Cy3 (red fluorescence).

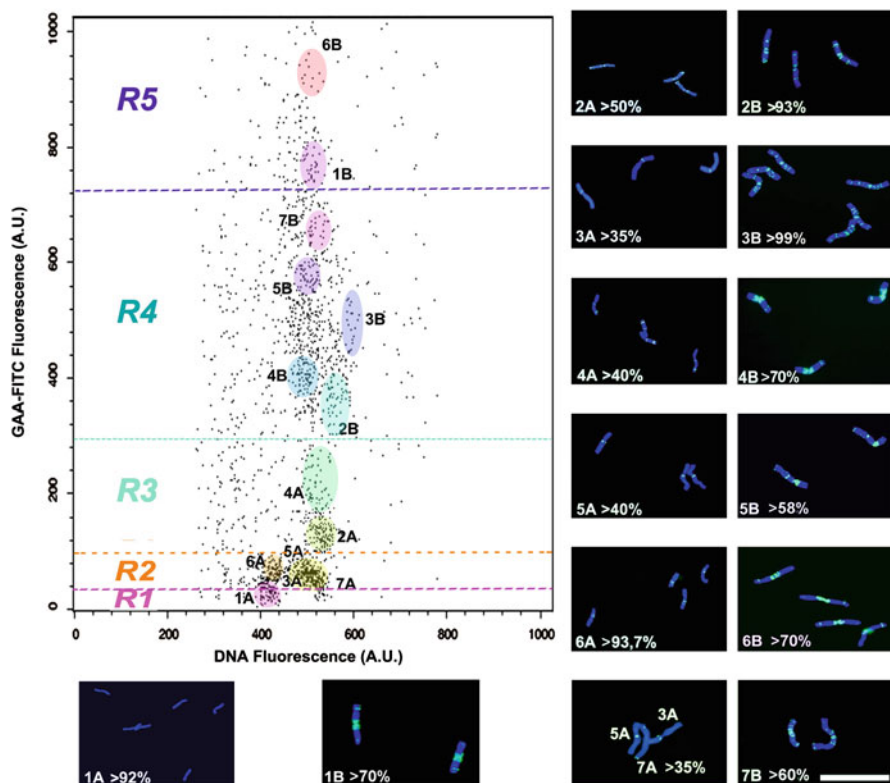


Fig. 13.8 Biparametric flow karyotyping and sorting gates of FISHIS (GAA)₇-FITC labeled *T. durum* Creso chromosomes. Wheat chromosomes were simultaneously stained for DNA content (DAPI, blue signal) and labeled by FISHIS with (GAA)₇-FITC (green signal). The chromosomes exhibiting similar fluorescence intensities accumulated in close regions, as showed up by light colors. Sorting gates were drawn around each single region and corresponding particles were flow sorted for a complete identification. Separate panels show the sorted chromosomes, their FISHIS pattern, and purity in respect to the whole number of sorted chromosomes for that region (100 chromosomes were counted per region). The intensity of FISHIS labeling generated a proportional distribution of chromosomes, with the A-genome complement all within regions R1-R3 and the B-genome chromosomes included within region R4-R5. Regions R1 to R5 were used to calculate the MESF (Molecules of Equivalent Soluble Fluorescein) values, to demonstrate the proportionality of FISHIS labeling in respect to the total band fluorescence intensities per chromosome, and to calculate the number of molecules of fluorescein incorporated by the different chromosomes (Giorgi et al. 2013). Bar = 10 μ m. (From doi: doi:10.1371/journal.pone.0057994.g002)

13.3.2 An Open Access to Potentially All Chromosomes in Triticeae

FISHIS proved to be effective in labeling chromosomes in several cereals (Table 13.2). By this technique, single type chromosome fractions were flow-sorted from each line of interest, regardless of the chromosome relative size or the existence

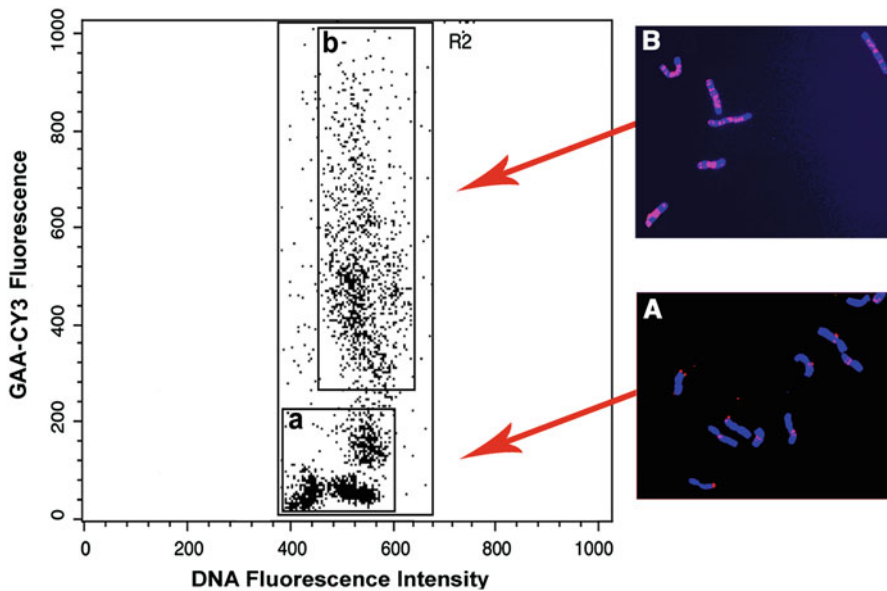


Fig. 13.9 Flow sorting of FISHIS labeled chromosome belong to the A- and B-genome. *T. durum* Creso chromosomes were FISHIS labeled with (GAA)₇-Cy3 and the resulting biparametric dot plot was divided by two regions, **a** and **b**, each containing all the chromosomes from A-genome and B-genome, respectively. The GAA banding pattern provided an excellent separation between the two groups and for the first time the homeologous genomes were flow sorted with a purity of more than 98 %

of cytogenetic mutants. We focus our description on three case studies, such as *T. durum*, *T. aestivum* and *D. villosum*.

In *T. durum* (pasta wheat), the standard flow karyotyping based on DAPI staining showed three main peaks, only one of which contained a single-type chromosome, that is 3B (Fig. 13.3).

FISHIS labeling combining DAPI and (GAA)₇-FITC fluorescence allows to discriminate some chromosome clusters and single-type fractions (Fig. 13.8).

Less intensely labeled A-genome chromosomes were assigned at the regions of lower (GAA)₇-FITC fluorescence intensity in both mono- and bi-parametric flow karyograms (Fig. 13.8: regions R1, R2, R3). The B-genome chromosomes, all with a high intensity and composite (GAA)₇ hybridization pattern, were recovered in regions related to higher levels of fluorescence within the karyograms (Fig. 13.8: regions R4, R5). Single-chromosome fractions were also located on karyograms allowing for the flow sorting of chromosome 1A to a purity of > 92 %, chromosome 6A to > 93 % purity, chromosome 2B to > 93 % purity, and chromosome 3B to 99 % purity.

As could be predicted by the GAA-FISH labeling pattern, the clear differences in GAA content among the A- and B-genome are also clearly demonstrated by FISHIS and allow for an easy separation of the A- and B-genome chromosomes (Fig. 13.9).

In *T. aestivum*, the flow karyotyping of DAPI-stained chromosomes in suspension raised four main peaks, with chromosome 3B the only one which could be

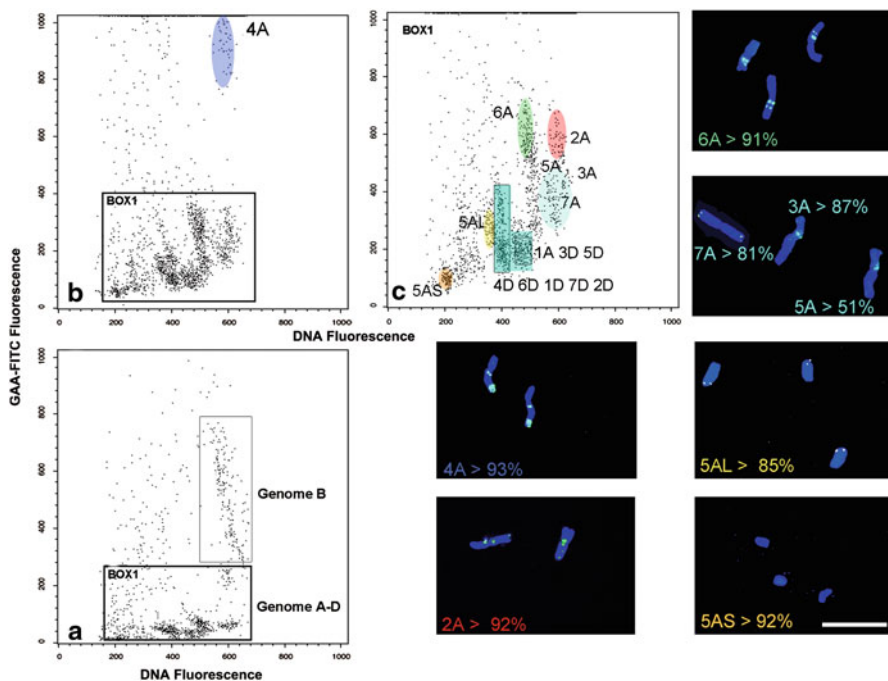


Fig. 13.10 FISHIS-based flow sorting of bread wheat chromosomes (*T. aestivum* Chinese Spring double ditelosomic line dDt5A) labeled by (GAA)₇-FITC. **a** Biparametric dot plot of chromosomes DNA content versus FISHIS (GAA)₇-FITC labeling. In BOX1 the homeologous A- and D-genome are presented, while the B-genome shows higher FISHIS fluorescence intensities; **b** increasing of the instrumental sensitivity allows a better discrimination of the chromosome 4A (colored region); **c** BOX1 shows all the D-genome chromosomes with chromosome 1A as the only contamination from the homeologous A-genome. Chromosomes 2A, 6A, and chromosome arms 5AS and 5AL can be sorted at high purity level. Bar = 10 μ m. (From doi: doi:10.1371/journal.pone.0057994.g004)

discriminated (Kubaláková et al. 2002). FISHIS labeling, which we applied to cv. Provinciale and line CSdDt5A (a Chinese Spring cytogenetic stock previously used for the isolation and sequencing of chromosome 5A arms; Vitulo et al. 2011), made it possible to differentiate between all three homeologous genomes, with the only contamination of 1A into the D-genome. A further discrimination inside each genome region where single-type chromosomes (e.g. 2A, 4A and 6A) were located is also shown in Fig. 13.10.

FISHIS labeling generated a complex pattern of fluorescent intensities which could be exclusively described by a high dynamic range of fluorescent values. By modulating the instrument responsiveness we optimized the discrimination among several clusters of fluorescence which were assigned to specific chromosomes and chromosome groups. In Fig. 13.10 we showed how the B-genome of bread wheat can be entirely located within the upper region of the dot plot and chromosome 4A could be flow-sorted to a purity of > 93% (Fig. 13.10b, region in color). The D-genome could also be discriminated from the A-genome on the basis of a weak GAA labeling

pattern and a slightly higher DNA content, except for chromosome 1A (Fig. 13.10c). Two more chromosomes could be sorted at the high purity of > 92 and > 91 %, as chromosomes 2A and 6A, respectively. To validate each flow sorting, chromosome purity checking can be done immediately, on FISHIS labeled samples allowing for an easy assessment of the sorted fraction composition (see telocentrics 5AS and 5AL: Fig. 13.10c).

Additional fluorescent oligos, other than the GAA probe, proved to generate a labeling pattern able to discriminate and hence enabling flow sorting of the wheat chromosomes (Table 13.2). The (AG)₁₂-Cy3 microsatellite showed four hybridization signals onto both pasta and bread wheat 3B, 4B, 5B and 6B chromosomes on slide, allowing the flow sorting of chromosome 5B (double strong band) and 3B to a purity level > 90 % (Fig. 13.11a). The pTa71 probe, known to specifically label rDNA at the satellite region, effectively labeled chromosomes 1B and 6B in FISHIS-labeled chromosome suspensions of *T. aestivum* dDt5A CS, allowing for the flow sorting of both chromosomes together (Fig. 13.11b).

In *D. villosum* (mosquitograss), a cross-compatible wild relative of wheat (Greadzielewska 2006) we found that (GAA)₇ can discriminate all of the seven chromosomes (VV genome) of the complement (Fig. 13.1; Grosso et al. 2012). Since this is a not common situation in FISH labeling, we have utilized this labeling pattern as an opportunity to developed a flow FISHIS karyotyping which characterized and allowed sorting of all the *D. villosum* chromosomes as single-type fractions. DAPI stained mosquito grass chromosomes presented a composite flow karyotyping made of four undistinguishable peaks except for the one at higher fluorescent value containing chromosome 6V (Fig. 13.12a). FISHIS labeling with (GAA)₇-Cy3 generated an intensity pattern, which combined with the fluorescence related to the DNA chromosome content after DAPI staining, allowed for the first whole single-type chromosome flow sorting from an eukaryote using a molecular “feature” of its DNA. All *D. villosum* chromosomes were flow-sorted at high purity measuring from 80–95 % (Fig. 13.12b, colored regions).

13.3.3 *The Bad and the Good: Limitations and Perspectives*

Many attempts have been carried out to find a consistent way to combine the analytical power of FISH and the quantitative flow sorting to characterize and isolate FISH-labeled chromosomes in suspension, but many glitches, such as chromosome stickiness and fragility, unreliable hybridization patterns and loss of chromosome shape, prevented reaching the target (Vandekken et al. 1990; Pich et al. 1995; He et al. 2001; Steinhäuser et al. 2002; Ma et al. 2005; Robertson and Thach 2009; Doležel et al. 2012). FISHIS overcomes all these limitations by using a simple washless method (Lucretti et al. 1999) with alkali DNA denaturation and *in situ* measurable hybridization of plant chromosomes in suspension using fluorescence-labeled DNA probes (Fig. 13.13). The consistency of labeling, a “weak point” in most of the previous experimental approaches to FISH in suspension, has been assessed by measuring the amount of fluorescence emitted by the FISHIS-labeled chromosomes in respect

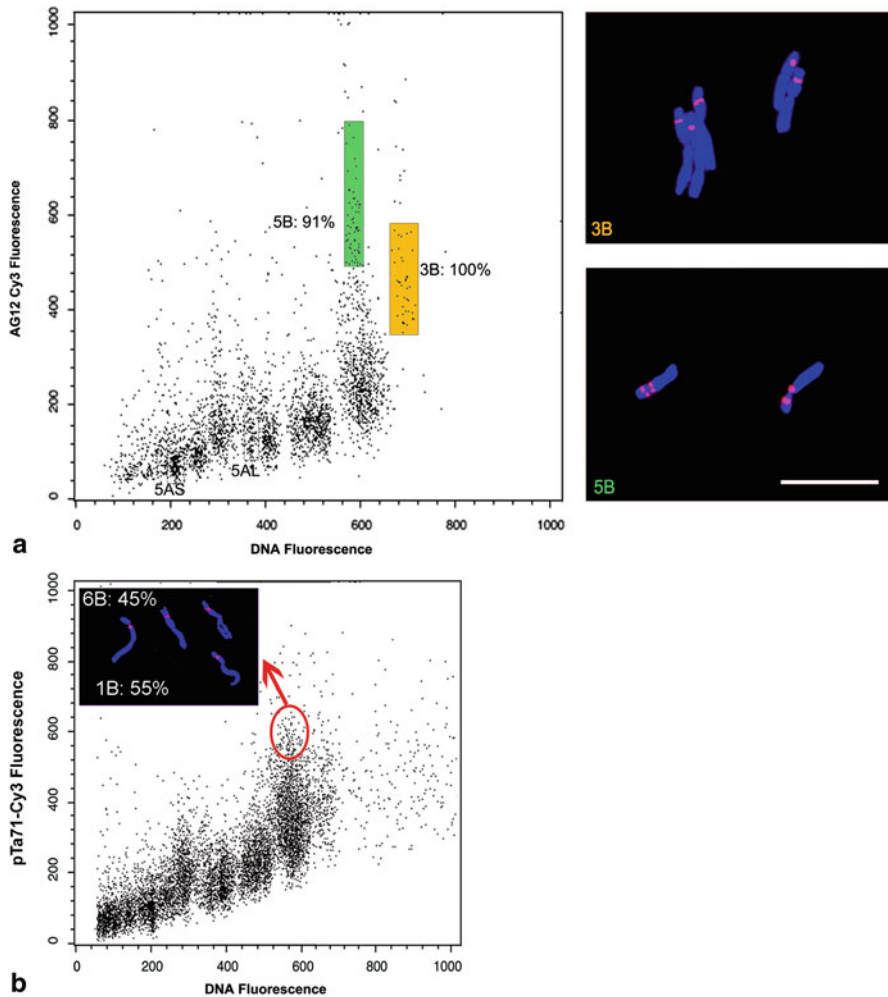


Fig. 13.11 Flow sorting of FISHIS labeled chromosomes with the oligo (AG)₁₂-Cy3 and the DNA probe pTa71-Cy3. Two separate sorting were performed using the FISHIS (AG)₁₂-Cy3 and pTa71-Cy3 flow karyotyping. In **a** sorting regions were drawn on a DAPI DNA content versus a (AG)₁₂-Cy3 FISHIS fluorescence dot plot in *T. aestivum* cv Chinese Spring double ditelosomic line dDt5A, which enclosed chromosomes 5B and 3B at a very high purity. In **b**, FISHIS chromosomes of *T. durum* Creso were flow sorted after pTa71-Cy3 labeling: the sorting region drawn on the dot plot allowed the isolation of both the satellite chromosomes 1B and 6B, which for the first time were discriminated from all the others. Bar = 10 μm. (Modified from doi: doi:10.1371/journal.pone.0057994.g005)

to a reference standard made with a known amount of fluorescein, and demonstrating a direct correspondence between the number of bands and the amount of fluorescence intensity (Giorgi et al. 2013; Fig. 13.13).

Labeled chromosomes can than be flow sorted enclosing wanted particles into a sorting gate, and their purity can be checked in real-time under a fluorescence

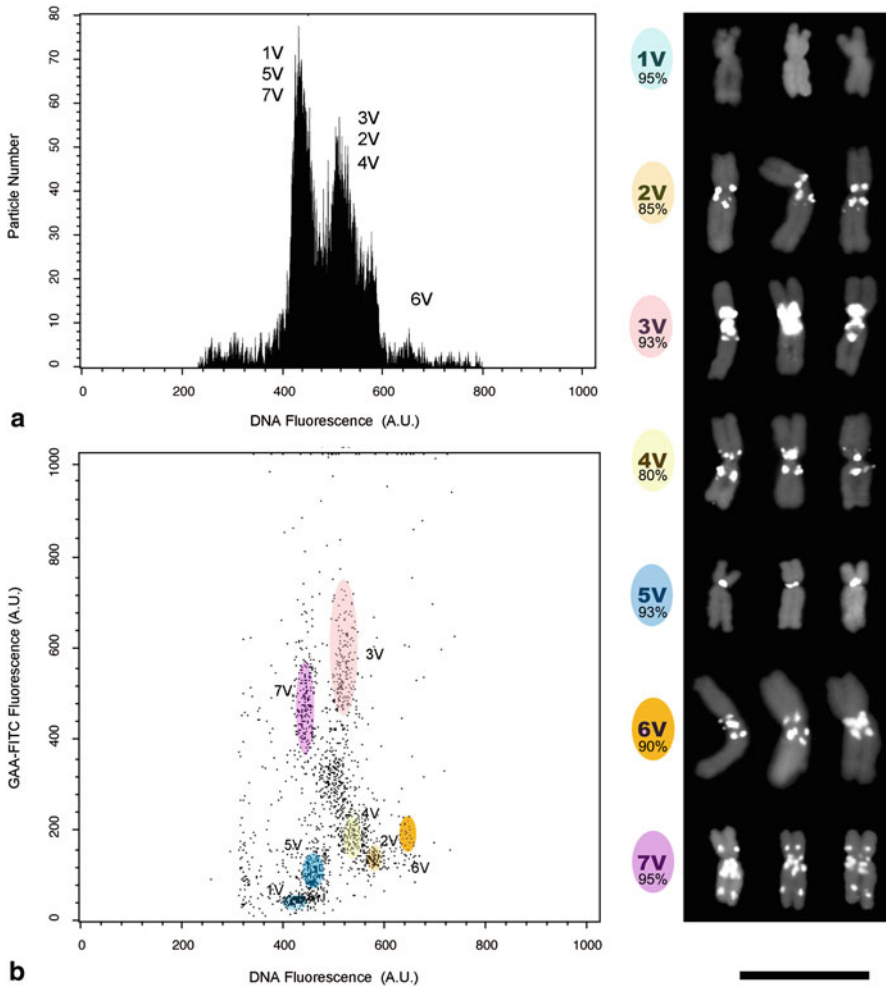


Fig. 13.12 Flow karyotyping and sorting of each chromosome of *D. villosum* complement on the bases of DNA content and non random distribution of the repetitive DNA sequences GAA. **a** A standard DNA content flow karyotyping resolves chromosomes 6V only out of the seven which compose the *D. villosum* complement. **b** a dot plot analysis of DNA content versus FISHIS (GAA)₇-FITC labeled chromosomes can discriminate all the seven chromosomes (colored regions) which can then be flow-sorted to a purity of 80–95% (specific purity percentage in Panels). Bar = 10 µm. (From doi: doi:10.1371/journal.pone.0057994.g006)

microscope, since all the particles of interest would have their own “sign of distinctiveness”. This feature will render an easy task to assess the purity of sorted samples. This is an important side-achievement of FISHIS, since purity is one of the “most desired” features in flow genomics, as the presence of a variable percentage of contaminating chromosomes would make the difference among a sample easy to assemble and a “fuzzy” one. A contamination of 2–15% in sorted samples has been


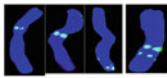

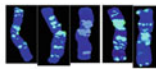

Chromosome Cluster	Chromosome type	Number of FISHIS Bands	MESF Values (average)	Chromosome pattern
R1	1A	0	4,971	
R2	3A 5A7A 6A	1-2	9,122	
R3	4A 2A	3-4	16,872	
R4	2B 3B 4B 5B 7B	7-9	39,405	
R5	1B 6B	9-10	140,192	

Fig. 13.13 FISHIS labeling intensities are proportional to the number and size of the banding pattern. The analysis of the FITC fluorescence intensities from sorted chromosomes contained in each of the regions marked as R1 to R5 in Fig. 13.8, showed a direct correlation among number and intensity of bands and their total fluorescence emission. The small band shown on chromosome 3A (Fig. 13.7) was selected as an arbitrary reference unit for bands number and intensity estimation. The number of FITC molecules incorporated in each chromosome banding pattern was calculated in respect to a fluorescence standard (Quantum FITC-5 MESF kit, BangsLab, IN, USA). FITC median fluorescence intensities were converted to an absolute unit of fluorescence as Molecules of Equivalent Soluble Fluorochrome (MESF). MESF values allow to measure the instrument sensitivity, to compare data among flow cytometers and to calculate FISHIS efficiency in terms of the amount of molecules of fluorescein attached to the sample. (From doi: doi:10.1371/journal.pone.0057994.s005)

reported, and this is a shared feature of flow sorters (Vitulo et al. 2011; Doležel et al. 2012). In sorted fractions, this level of impurity did not jeopardized NGS results (Šimková et al. 2008; Vitulo et al. 2011). FISHIS is providing flow cytogenomics with a new type of samples because it increases the quality of discrimination of chromosomes in suspension, adding a molecular labeling to the basic feature of their DNA content. The first successful flow sorting of the whole chromosome set of an eukaryote with alike chromosomes has been achieved for the plant *D. villosum* (Giorgi et al. 2013) using a microsatellite DNA probe generating a specific chromosome hybridization pattern (Grosso et al. 2012), which the flow cytometer can distinguish not as a distribution banding pattern but as fluorescence intensity changes. The quality of DNA obtained after FISHIS labeling and flow sorting, substantiated its use for molecular manipulations such as direct PCR, Multiple Displacement Amplification (MDA) (Giorgi et al. 2013) and the construction of NGS sequencing libraries (Lucretti et al.: work in progress).

The construction of genetic maps will benefit from the availability of single chromosome types from elite varieties and wild species which can be used for direct physical localization of a specific sequence by PCR amplification of DNA samples

made of few 100 sorted chromosomes. Either, few nanograms of sorted DNA can be augmented to micrograms by MDA (Šimková et al. 2008) for the generation of thousands of chromosome-specific molecular markers saturating the whole chromosome allowing a deep understanding of the gene content and organization (Wenzl et al. 2010; Nie et al. 2012).

New insight in plant polyploidy could be obtained by the novel possibility of separating homeologous genomes from the same nucleus, as demonstrated in *T. durum* Creso where FISHIS-labeled A- and B-genome chromosomes can be independently flow sorted at a high purity (Giorgi et al. 2013). This holds true for *T. aestivum* as well, where the D-genome can be isolated from the A- and B-genomes, with a single contamination from only chromosome 1A.

Microsatellite motifs and rDNA sequences are diffused in a non-random chromosome distribution allowing unequivocal chromosome identification (Kalia et al. 2011; Cuadrado and Jouve 2010). It is foreseeable that this technique would also help in analyzing and sorting animal chromosomes, with peculiar cytogenetic abnormalities such as copy number variations and other cytogenetic abnormalities which can be revealed by a repetitive DNA probe or by other kinds of DNA probes such as rDNA, or possibly exogenous transfecting DNA sequences (Choy et al. 2010).

Chromosome sorting based on FISHIS specific hybridization pattern is an innovative approach which can help in developing new genetic resources from every plant species where DNA probes such as high repetitive DNA sequences are available, and high quality chromosome suspensions could be obtained.

The use of such handy and inexpensive DNA probes combined with the easy FISHIS technique would allow for the sorting of almost any chromosomes and nuclei, a great benefit fact for any *de novo* sequencing and re-sequencing project in germplasm collections and plant genomics.

References

- Ageno M, Dore E, Frontali C (1969) The alkaline denaturation of DNA. *Biophys J* 9:1281–1311
- Arumuganathan K, Slattey JP, Tanksley SD, Earle ED (1991) Preparation and flow cytometric analysis of metaphase chromosomes of tomato. *Theor Appl Gen* 82:101–111
- Bedbrook J, Jones JDG, O'Dell M, Thompson RD, Flavell RB (1980) Molecular description of telomeric heterochromatin in *Secale* species. *Cell* 19:545–560
- Bento M, Gustafson JP, Viegas W, Silva M (2011) Size matters in Triticeae polyploids: larger genomes have higher remodeling. *Genome* 54:175–183
- Bradbury L, Niehaus T, Hanson A (2013) Comparative genomics approaches to understanding and manipulating plant metabolism. *Curr Opin Biotech* 24:278–284
- Brind'Amour J, Lansdorp PM (2011) Analysis of repetitive DNA in chromosomes by flow cytometry. *Nat Methods* 8:484–486
- Carmona Á, Friero E, de Bustos A et al (2013) Cytogenetic diversity of SSR motifs within and between *Hordeum* species carrying the H genome: *H. vulgare* L. and *H. bulbosum* L. *Theor Appl Genet* 126:949–961
- Chen J, Huang Q, Gao D et al (2013) Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nat Commun* 4:1595

- Chester M, Leitch AR, Soltis PS, Soltis DE (2010) Review of the application of modern cytogenetic methods (FISH/GISH) to the study of reticulation (Polyploidy/Hybridisation). *Genes* 1:166–192
- Choy KW, Setlur SR, Lee C, Lau TK (2010) The impact of human copy number variation on a new era of genetic testing. *BJOG* 117:391–398
- Claros MG, Bautista R, Guerrero-Fernández D et al (2012) Why assembling plant genome sequences is so challenging. *Biology* 1:439–459
- Comai L (2005) The advantages and disadvantages of being polyploid. *Nat Rev Genet* 6:836–846
- Conia J, Bergounioux C, Perennes C et al (1987) Flow cytometric analysis and sorting of plant chromosomes from *Petunia hybrida* protoplasts. *Cytometry* 8:500–508
- Consortium TPGS (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–195
- Cook DR, Varshney RK (2010) From genome studies to agricultural biotechnology: closing the gap between basic plant science and applied agriculture. *Curr Opin Plant Biol* 13:115–118
- Cuadrado Á, Schwarzacher T (1998) The chromosomal organization of simple sequence repeats in wheat and rye genomes. *Chromosoma* 107:587–594
- Cuadrado Á, Schwarzacher T, Jouve N (2000) Identification of different chromatin classes in wheat using in situ hybridization with simple sequence repeat oligonucleotides. *Theor Appl Genet* 101:711–717
- Cuadrado Á, Jouve N (2007) The nonrandom distribution of long clusters of all possible classes of trinucleotide repeats in barley chromosomes. *Chromosome Res* 15:711–720
- Cuadrado Á, Cardoso M, Jouve N (2008a) Increasing the physical markers of wheat chromosomes using SSRs as FISH probes. *Genome* 51:809–815
- Cuadrado Á, Cardoso M, Jouve N (2008b) Physical organisation of simple sequence repeats (SSRs) in *Triticeae*: structural, functional and evolutionary implications. *Cytogenet Genome Res* 120:210–219
- Cuadrado Á, Golczyk H, Jouve N (2009) A novel, simple and rapid nondenaturing FISH (ND-FISH) technique for the detection of plant telomeres. Potential used and possible target structures detected. *Chromosome Res* 17:755–762
- Cuadrado Á, Jouve N (2010) Chromosomal detection of simple sequence repeats (SSRs) using nondenaturing FISH (ND-FISH). *Chromosoma* 19:495–503
- DeBolt S (2010) Copy number variation shapes genome diversity in *Arabidopsis* over immediate family generational scales. *Genome Biol Evol* 2:441–453
- Delaat A, Blaas J (1984) Flow-cytometric characterization and sorting of plant chromosomes. *Theor Appl Gen* 67:463–467
- Doležel J, Čiháľková J, Lucretti S (1992) A high-yield procedure for isolation of metaphase chromosomes from root-tips of *Vicia faba* L. *Planta* 188:93–98
- Doležel J, Čiháľková J, Weiserova J, Lucretti S (1999) Cell cycle synchronization in plant root meristems. *Methods Cell Sci* 21:95–107
- Doležel J, Kubaláková M, Paux E et al (2007) Chromosome-based genomics in the cereals. *Chromosome Res* 15:51–66
- Doležel J, Šimková H, Kubaláková M et al (2009) Chromosome genomics in the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) *Genetics and genomics of the Triticeae*. Springer Science+Business Media, Heidelberg, pp 385–316
- Doležel J, Vrána J, Šafář J et al (2012) Chromosomes in the flow to simplify genome analysis. *Funct Integr Genomic* 12:397–416
- Edwards D, Batley J, Snowdon RJ (2013) Accessing complex crop genomes with next-generation sequencing. *Theor Appl Gen* 126:1–11
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16:77–88
- Galbraith DW (2010) Flow cytometry and fluorescence-activated cell sorting in plants: the past, present, and future. *Biomédica* 30:65–70.
- Galbraith DW, Lucretti S (2000) Large particle sorting. In: Radbruch A (ed) *Flow cytometry and cell sorting*. Springer Berlin Heidelberg, Heidelberg, pp 293–317

- Gerlach W, Bedbrook J (1979) Sequence organization of the repeating units in the nucleus of wheat which contain 5s ribosomal-RNA genes. *Nucleic Acids Res* 8:4851–4865
- Gill KS, Arumuganathan K, Lee JH (1999) Isolating individual wheat (*Triticum aestivum*) chromosome arms by flow cytometric analysis of ditelosomic lines. *Theor Appl Gen* 98:1248–1252
- Giorgi D, Farina A, Grosso V et al (2013) FISHIS: Fluorescence *In situ* hybridization in suspension and chromosome flow sorting made easy. *PLoS ONE* 8:e57994
- Greadzielewska A (2006) The genus *Dasyphyrum*—part 2. *Dasyphyrum villosum*—a wild species used in wheat improvement. *Euphytica* 152:441–454
- Greilhuber J (2005) Intraspecific variation in genome size in angiosperms: identifying its existence. *Ann Bot* 95:91–98
- Grosso V, Farina A, Gennaro A et al (2012) Flow sorting and molecular cytogenetic identification of individual chromosomes of *Dasyphyrum villosum* L. (*H. villosa*) by a Single DNA Probe. *PLoS ONE* 7:e50151
- Gualberti G, Doležel J, Macas J, Lucretti S (1996) Preparation of pea (*Pisum sativum* L.) chromosome and nucleus suspensions from single root tips. *Theor Appl Gen* 92:744–751
- Hansey CN, Vaillancourt B, Sekhon RS et al (2012) Maize (*Zea mays* L.) Genome Diversity as Revealed by RNA-Sequencing. *PLoS ONE* 7:e33071
- He H, Deng W, Cassel MJ, Lucas JN (2001) Fluorescence *in situ* hybridization of metaphase chromosomes in suspension. *Int J Radiat Biol* 77:787–795
- Henson J, Tischler G, Ning ZM (2012) Next-generation sequencing and large genome assemblies. *Pharmacogenomics* 13:901–915
- Heslop-Harrison J, Schwarzacher T (2011) Organization of the plant genome in chromosomes. *Plant J* 66:18–33
- Jiang J, Gill BS (2006) Current status and the future of fluorescence *in situ* hybridization (FISH) in plant genome research. *Genome* 49:1057–1068
- Kalia RK, Rai MK, Kalia S et al (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica*, 309–334
- Kato A, Vega JM, Han F et al (2005) Advances in plant chromosome identification and cytogenetic techniques. *Curr Opin Plant Biol* 8:148–154
- Kaul S, Koo HL, Jenkins J et al (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Kubaláková M, Vrána J, Čihalíková J et al (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* 104:1362–1372
- Kubaláková M, Valarik M, Bartoš J et al (2003) Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome* 46:893–905
- Kubaláková M, Kovářová P, Suchánková P et al (2005) Chromosome Sorting in Tetraploid Wheat and Its Potential for Genome Analysis. *Genetics* 170:823–829
- Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Lee J-H, Ma Y, Wako T et al (2004) Flow karyotypes and chromosomal DNA contents of genus *Triticum* species and rye (*Secale cereale*). *Chromosome Res* 12:93–102
- Li W, Chen P, Qi L, Liu D (1995) Isolation, characterization and application of a species-specific repeated sequence from *Haynaldia villosa*. *Theor Appl Genet* 90:526–533
- Lucretti S, Doležel J (1995) Cell cycle synchronization, chromosome isolation, and flow-sorting in plants. *Methods Cell Biol* 50:61–83
- Lucretti S, Doležel J, Schubert I, Fuchs J (1993) Flow karyotyping and sorting of *Vicia faba* chromosomes. *Theor Appl Gen* 85:665–672
- Ma YZ, Lee JH, Li LC et al (2005) Fluorescent labeling of plant chromosomes in suspension by FISH. *Genes Genet Syst* 80:35–39
- Macas J, Doležel J, Gualberti G et al (1995) Primer-induced labeling of pea and field bean chromosomes *in situ* and in suspension. *Biotechniques* 19:402–404; 407–408
- Macas J, Doležel J, Lucretti S et al (1993) Localization of seed protein genes on flow-sorted field bean chromosomes. *Chromosome Res* 1:107–115

- Mackay J, Dean JD, Plomion C et al (2012) Towards decoding the conifer giga-genome. *Plant Mol Biol* 80:555–569
- Metzker ML (2010) Sequencing technologies: the next generation. *Nat Rev Genet* 11:31–46
- Mizuno H, Wu J, Katayose Y et al (2008) Chromosome-specific distribution of nucleotide substitutions in telomeric repeats of rice (*Oryza sativa* L.). *Mol Biol Evol* 25:62–68
- Molnár I, Kubaláková M, Simkova H et al (2011) Chromosome isolation by flow sorting in *Aegilops umbellulata* and *Ae. comosa* and their allotetraploid hybrids *Ae. biuncialis* and *Ae. geniculata*. *PLoS One* 6:e27708
- Morgante M, De Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. *Curr Opin Plant Biol* 10:149–155
- Nie X, Li B, Wang L et al (2012) Development of chromosome-arm-specific microsatellite markers in *Triticum aestivum* (*Poaceae*) using NGS technology. *Am J Bot* 99:e369–e371
- Page JT, Gingle AR, Udall JA (2013) PolyCat: a resource for genome categorization of sequencing reads from allopolyploid organisms. *G3* 3:517–525
- Paterson AH, Bowers JE, Burrow MD et al (2000) Comparative genomics of plant chromosomes. *Plant Cell* 12:1523–1539
- Payseur BA, Jing P, Haas RJ (2011) A genomic portrait of human microsatellite variation. *Mol Biol Evol* 28:303–312
- Pedersen C, Langridge P (1997) Identification of the entire chromosome complement of bread wheat by two-colour FISH. *Genome* 40:589–593
- Pellicer J, Fay MF, Leitch IJ (2010) The largest eukaryotic genome of them all? *Bot J Lin Soc* 164:10–15
- Pich U, Meister A, Macas J et al (1995) Primed *in-situ* labeling facilitates flow sorting of similar sized chromosomes. *Plant J* 7:1039–1044
- Pruitt RE, Meyerowitz EM (1986) Characterization of the genome of *Arabidopsis thaliana*. *J Mol Biol* 187:169–183
- Raap AK, Marijnjen JG, Vrolijk J, Ploeg M van der (1986) Denaturation, renaturation, and loss of DNA during *in situ* hybridization procedures. *Cytometry* 7:235–242
- Robertson KL, Thach DC (2009) LNA flow FISH: a flow cytometry fluorescence *in situ* hybridization method to detect messenger RNA using locked nucleic acid probes. *Anal Biochem* 390:109
- Šafař J, Bartoš J, Janda J et al (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J* 39:960–968
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74:5463–5467
- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Sears E, Sears L (1978) The telocentric chromosomes of common wheat. In: Ramanujam S (ed) *Indian Agricultural Research Institute, New Delhi*, pp 389–407
- Sharma S, Raina SN (2005) Organization and evolution of highly repeated satellite DNA sequence in plant chromosome. *Cyt Gen Res* 109:15–26
- Šimková H, Svensson JT, Condamine P et al (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* 9:294
- Šmarda P, Bureš P, Horová L, Rotreklová O (2008) Intrapopulation genome size dynamics in *Festuca pallens*. *Ann Botany* 102:599–607
- Smith DB, Flavell RB (1975) Characterisation of wheat genome by renaturation kinetics. *Chromosoma* 50:223–242
- Speel EJM (1999) Detection and amplification systems for sensitive, multiple-target DNA and RNA *in situ* hybridization: Looking inside cells with a spectrum of colors. *Hist Cell Biol* 112:89–113
- Steinhaeuser U, Starke H, Nietzel A et al (2002) Suspension (S)-FISH, a new technique for interphase nuclei. *J Histochem Cytochem* 50:1697–1698
- Sutton T, Baumann U, Hayes J et al (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* 318:1446–1449

- Trask B, Vandenberg G, Landegren J et al (1985) Detection of dna-sequences in nuclei in suspension by *in situ* hybridization and dual beam flow-cytometry. *Science* 230:1401–1403
- Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Genetics* 13:36–46
- Vandekken H, Arkesteijn GJA, Visser JWM, Bauman JGJ (1990) Flow cytometric quantification of human-chromosome specific repetitive dna-sequences by single and bicolor fluorescent insitu hybridization to lymphocyte interphase nuclei. *Cytometry* 11:153–164
- Velasco R, Zharkikh A, Troggo M et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *Plos One* 2:e1326
- Vitulo N, Albiero A, Forcato C et al (2011) First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *Plos One* 6:e26421
- Vrána J, Kubaláková M, Simkova H et al (2000) Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics* 156:2033–2041
- Wang XH, Lazzeri PA, Lorz H (1992) Chromosomal variation in dividing protoplasts derived from cell-suspensions of barley (*Hordeum vulgare* L). *Theor Appl Gen* 85:181–185
- Wang Y, Wang X, Paterson AH (2012) Genome and gene duplications and gene expression divergence: a view from plants. *Ann NY Acad Sci* 1256:1–14
- Wenzl P, Suchankova P, Carling J et al (2010) Isolated chromosomes as a new and efficient source of DArT markers for the saturation of genetic maps. *Theor Appl Gen* 121:465–474
- Xu K, Xu X, Fukao T et al (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 442:705–708

Chapter 14

Mining Genetic Resources *via* Ecotilling

Bradley J. Till

Contents

14.1 Introduction	350
14.2 Origins of Ecotilling	350
14.3 Ecotilling the Plant Kingdom	353
14.3.1 Germplasm Characterization	355
14.3.2 Ecotilling for Functional Genomics	357
14.4 Polymorphism Discovery Methods Adapted for Ecotilling	359
14.5 Future Directions	362
References	363

Abstract The acquisition of genomic sequence data from plants has grown dramatically in the past decade. Sequence information from a single individual provides important insight into gene function, genome architecture and evolutionary relationships between species.

Methods for rapid discovery and characterization of nucleotide polymorphisms have enabled investigations into nucleotide variation in a population of individuals of the same species, allowing deeper insight into genetic variation, gene function and flow in germplasm collections and wild populations. The Ecotilling method was developed as a high-throughput and low cost platform for the discovery of SNPs and small indels. Since its inception, Ecotilling has been adapted for more than 20 plant species for a range of applications including population genetics, mapping, and QTL cloning. I review here progress in the establishment of Ecotilling for diploid and polyploid plants, adaptation of the method for a variety of different investigations, and modifications to mutation discovery methods.

B. J. Till (✉)
Plant Breeding and Genetics Laboratory,
Joint FAO/IAEA Division of Nuclear Techniques
in Food and Agriculture, International Atomic Energy Agency,
Vienna International Centre,
P.O. Box 100, 1400 Vienna, Austria
e-mail: b.till@iaea.org

Keywords Polymorphism discovery · Enzymatic mismatch cleavage · CEL I · Celery Juice Extract · SNP · Indel ·

Abbreviations

SNP	Single Nucleotide Polymorphism
Indel	A nucleotide insertion or deletion
QTL	Quantitative Trait Locus
TILLING	Targeting Induced Local Lesions IN Genomes
CJE	Crude Celery Juice Extract containing CEL I nuclease
dHPLC	denaturing High Pressure Liquid Chromatography
bp	Base pairs
kb	kilobase pairs

14.1 Introduction

Heritable nucleotide variation is a major contributor to intra and inter-species phenotypic variation. It was this resource that was exploited by the first farmers over 10,000 years ago for the domestication of crops. In the present day, natural and induced nucleotide variation continues to be a rich resource for the plant breeder, evolutionary biologist and basic researcher studying gene function. Historically, technological developments have allowed researchers to enter into new avenues of research and have driven new discoveries. Genomics is no exception. Sequencing technologies became sufficiently advanced to allow the acquisition of whole genome assemblies from complex eukaryotes (Kaul et al. 2000). With the release of the first plant genome, *Arabidopsis thaliana*, efforts have increased in the genomic evaluation of intra and interspecific variation (Caicedo and Purugganan 2005). Such evaluations can provide important insights into gene function, the diversity of germplasm collections, and the effects of selective pressures on populations. However, at the time of release of the *Arabidopsis* genome in 2000, cost effective methods didn't exist that would allow rapid and accurate surveys of nucleotide polymorphisms in standard research laboratories. Ecotilling methodologies were developed to bridge this gap.

14.2 Origins of Ecotilling

Ecotilling for the discovery and characterization of natural nucleotide polymorphisms was adapted from methods developed for the reverse-genetics strategy known as TILLING (Targeting Induced Local Lesions in Genomes). First described in 2000, TILLING combines traditional mutagenesis with high-throughput mutation discovery (McCallum et al. 2000b). The first large-scale TILLING project was developed in *Arabidopsis thaliana* using the chemical mutagen ethyl methanesulfonate (EMS). The goal was to develop a public service where users could request allelic series of

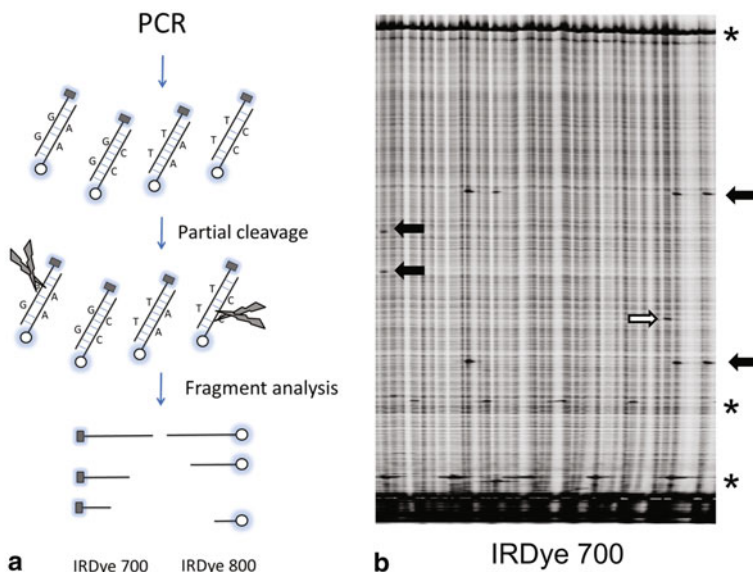


Fig. 14.1 Enzymatic mismatch cleavage and fluorescence detection of natural polymorphisms for Ecotilling. Target amplicons of ~750–1600 bp are typically amplified using gene-specific primers that are end labelled with fluorescent dyes IRDye700 (*box*) and IRDye800 (*circle*) **a**. After PCR, samples are denatured and annealed to create heteroduplexed molecules that lack hydrogen bonding at polymorphic regions. These mismatches are the substrate for partial enzymatic cleavage using a single-strand specific nuclease. Partial cleavage allows for the generation of labelled DNA fragments for all polymorphic sites. Complete digestion would allow visualization of only terminal polymorphisms. **b**. An early example of Ecotilling from 2004. The IRDye700 image channel from 32 lanes of a 96 lane *Arabidopsis* TILLING Project gel is shown. 768 plants from a TILLING population were interrogated for induced mutations in a 1 kb gene fragment. In addition to a rare induced mutation (marked by a white arrow), multiple common polymorphisms were identified within amplicons (black arrows). It was later determined that these represented natural SNPs from other *Arabidopsis* accessions that had inadvertently entered the TILLING population. Asterisks mark 95 and 200 bp lane markers that are added to assist lane identification. This image was produced by Anthony Odden and kindly supplied by Steven Henikoff, Fred Hutchinson Cancer Research Center

point mutations in their target gene of choice. It was important, therefore, to establish a high-throughput, high accuracy, and low cost mutation discovery platform. An enzymatic mismatch cleavage method was developed whereby ~1–1.5 kb gene fragments were amplified using gene-specific PCR primers that were fluorescently labelled on their 5' ends with IRDye700 (forward) or IRDye800 (reverse) dyes. After PCR, amplicons are denatured and annealed to create heteroduplexed molecules that are single-stranded in the region of the mutation. These single-stranded regions of the double stranded duplex are the substrate for enzymatic cleavage using the CEL I nuclease or a crude extract containing CEL I (Oleykowski et al. 1998). After cleavage, samples are size fractionated on denaturing polyacrylamide gels and mutations observed by fluorescence detection of cleaved fragments using a LI-COR DNA analyser (Fig. 14.1) (Colbert et al. 2001; Till et al. 2006b). Throughput was increased by pooling genomic DNA samples together prior to PCR. Efficient recovery

of heterozygous mutations was shown when pooling DNAs from 8 mutant Arabidopsis plants, suggesting false negative errors did not increase due to sample pooling (Greene et al. 2003). Running gels with 96 sample lanes allowed the interrogation of 768 individuals for rare induced mutations in a single gel run.

It was during the initial scale-up of TILLING prior to opening a public service that a peculiar phenomenon was observed. Unique Single Nucleotide Polymorphism (SNP) mutations were typically detected at a frequency of 4 mutations per 768 samples. Because mutagenesis is random, mutations were expected to be discovered in different regions of the target gene, as observed by the generation of different molecular weight bands on the gel. This was mostly true; however, a common multiple band pattern was observed in a subset of samples (Fig. 14.1b). It was first hypothesized that the multiple banding patterns might have been created from induced mutations in different samples in the same pool of samples that happened to be combined by chance. Replication of these bands in different sample pools (gel lanes) could then be explained by sample contamination. This was based on the presumption that if CEL I cleaved PCR products were end-labelled, only the terminal fragments linked to the label would be observable on the gel image, and so at most only two fragments from a sample should ever be observable (Fig. 14.1a). This hypothesis proved to be incorrect as multiple bands could be observed when mutation discovery was performed on genomic DNA from a single plant. Thus it was determined that the single-strand specific nuclease did not cleave substrates to completion under the reaction conditions used. In other words, there exists a probability that any SNP, but not all SNPs on the heteroduplexed DNA molecule are cleaved. There are many heteroduplexed molecules in each sample assayed, and some molecules escape cleavage at terminal SNP sites, allowing the observation of multiple bands on the gel and a cataloguing of all polymorphisms in the gene target. Sanger sequencing was used to confirm that the questionable bands on the TILLING gel represented true SNPs and not artifactual fragments. Ultimately, it was determined that plants showing multiple heterozygous SNPs represented only a small percentage of the total population. Further investigation showed that these samples represented a unique accession different than the TILLING lines, and had entered into the mutant population accidentally. To avoid confusion, contaminant lines were removed from the TILLING population. At this point it became clear that the methods developed for TILLING could be adapted for discovery and genotyping of higher frequency natural nucleotide polymorphisms in populations.

To test the efficacy of this approach, a pilot project was undertaken where DNA from 196 Arabidopsis accessions, or ecotypes, were prepared, arrayed in 96 well microtiter plates, and subjected to the TILLING protocol. Primer pairs for five gene targets ranging between 809 and 1034 bp that were previously used for TILLING were selected for the study. To avoid confusion with genotype assignments, it was decided that samples would not be pooled before screening. This revealed a limitation in the methodology; only heteroduplexed molecules are cleaved, thus homozygous allelic differences in the tested germplasm collection would go undetected. To circumvent this issue, an equal amount of reference DNA from the sequenced Columbia genotype was added prior to PCR. Screening samples alone, and in combination with

the reference sequence, allowed unambiguous assignment of the zygosity of alleles compared to the reference. A key point to this strategy is the use of a reference that is largely homozygous (described below). The pilot screen was successful, and sequencing of polymorphisms discovered by enzymatic mismatch cleavage showed the method to be highly accurate. In addition to the recovery of single nucleotide polymorphisms such as the GC-AT transition changes that predominate in TILLING screens, a wide range of changes could be recovered including other SNPs (transition and transversion), small deletions, and variations in microsatellite repeat number. The method provided a rapid visual picture of nucleotide diversity in populations that was anchored to specific genomic regions. For example, band analysis revealed the site of a possible introgression in the PIF2 gene target in one of the samples analysed. The method also proved to be highly cost effective when compared to traditional Sanger sequencing. Samples could be grouped according to haplotype, and only one member of the haplotype group needed to be sequenced to catalogue nucleotide variation. Informatics load was also greatly reduced as the approximate location of nucleotide polymorphism was determined by band molecular weight and only those regions of the generated sequence trace needed to be analysed. Knowledge of the approximate location of the polymorphism also allowed sequence validation to be performed from a single direction using the same primer designed for Ecotilling, thus further streamlining the process. Because TILLING methodologies were applied to *Arabidopsis ecotypes*, the method was termed *Ecotilling* (Comai et al. 2004). Computational analysis tools developed for TILLING were easily adapted for Ecotilling, including gene target selection and primer design (CODDLe) (McCallum et al. 2000a), estimations of the effect of polymorphisms on gene function (SIFT, PARSesNP) (Ng and Henikoff 2003; Taylor and Greene 2003) and gel analysis software (GelBuddy) (Zerr and Henikoff 2005). While subsequent publications have used variations of capitalization and hyphenation of the term Ecotilling (EcoTILLING, Eco-TILLING), the form as it appeared in the original publication will be used in this review.

14.3 Ecotilling the Plant Kingdom

When considering progress in developing Ecotilling platforms, one can consider two broad categories of advancement: the development of technologies or strategies for improved polymorphism discovery, and the establishment of Ecotilling in a wide range of organisms. At the time of writing, over 25 Ecotilling projects have been described. While there have been some important advancements in humans and animals, the majority of applications have been in plants (Table 14.1). This may in part be due to the fact that Ecotilling was first developed in *Arabidopsis thaliana*, and thus the method is more familiar to the plant research community. Another contributing factor may be the increasing genomic sequence information collected for plants and the large number of crop species important for food, feed, fuel and industrial materials. Ecotilling has been used for mapping, candidate gene discovery, evaluation of nucleotide diversity, population genetics, and other applications.

Table 14.1 Examples of ecotilling projects in plants

Species	Samples screened	Gene targets	Mutation discovery ^a	Focus	Reference
<i>Arabidopsis thaliana</i>	192	5	CEL I/LI-COR	Germplasm characterization	(Comai et al. 2004)
<i>Arachis hypogaea</i>	30	1	CJE/LI-COR	Functional genomics/allergens	(Ramos et al. 2009)
Genus Brassica	117	6	CEL I/ LI-COR/UL	Functional genomics/oil quality	(Wang et al. 2010)
Genus Capsicum	233	4	cDNA/CJE/LI-COR	Functional genomics/virus resistance	(Ibiza et al. 2010)
Genus Cucumis	113	3	ENDO-1/LI-COR	Functional genomics/virus susceptibility	(Nieto et al. 2007)
<i>Glycine max</i>	25	5	CJE/agarose	Functional genomics/protein quality	(Jegadeesan et al. 2012)
<i>Helianthus annuus</i>	112	24	CJE/CE or dHPLC	Germplasm characterization	(Fusari et al. 2011)
<i>Hordeum vulgare</i>	36	6	CEL I/CE	Functional genomics/powdery mildew resistance	(Mejlhede et al. 2006)
<i>Hordeum vulgare</i>	292	1	CEL I/ LI-COR/UL	Functional genomics/photosynthesis and agronomic traits	(Xia et al. 2012)
<i>Monochoria vaginalis</i>	4	2	CEL I/ LI-COR/UL	Functional genomics/herbicide resistance in weeds	(Wang et al. 2007)
Genus Musa	80	14	CJE/LI-COR	Germplasm characterization	(Till et al. 2010)
Genus Oryza	48	~14	Surveyor/LI-COR/UL	Germplasm characterization/population genetics	(Rakshit et al. 2007)
<i>Oryza sativa</i>	48	6	CJE/LI-COR & agarose gel	Genetic mapping	(Raghavan et al. 2007)
<i>Oryza sativa</i>	45	87	CJE/agarose gel	Functional genomics/boron toxicity	(Ochiai et al. 2011)
<i>Oryza sativa</i>	375	2	CJE/agarose gel	Functional genomics/salinity tolerance	(Negrao et al. 2011)
<i>Oryza sativa</i>	95	190	CEL I/agarose	Functional genomics/drought tolerance	(Yu et al. 2012)
<i>Phaseolus vulgaris</i>	4	37	CJE/agarose & PAGE	Marker development/virus resistance	(Galeano et al. 2009)
Genus Picea	~176	60	CJE/LI-COR/UL	Genetic mapping	(Rungis et al. 2005)

Table 14.1 (continued)

Species	Samples screened	Gene targets	Mutation discovery ^a	Focus	Reference
<i>Populus nigra</i>	768	5	Illumina sequencing	Functional genomics/lignin biosynthesis	(Marroni et al. 2011)
<i>Populus trichocarpa</i>	41	9	CJE/LI-COR	Germplasm characterization/population genetics	(Gilchrist et al. 2006)
Genus <i>Saccharum</i>	2	5	CJE/Surveyor/CE	Technology development for mutation discovery in polyploids	(Cordeiro et al. 2006)
<i>Solanum tuberosum</i>	3	1	CJE/LI-COR	Germplasm characterization	(Elias et al. 2009)
<i>Triticum aestivum</i>	214	1	CJE/agarose gel	Functional genomics/vernalization	(Chen et al. 2011)
Genus <i>Vigna</i>	22	10	CEL I/LI-COR	Germplasm characterization	(Barkley and Wang 2008)

^aUL Universal labelling with fluorescent dyes, CE capillary electrophoresis

14.3.1 Germplasm Characterization

Shortly after the first published description, Ecotilling was quickly adopted by other groups for the analysis of plant, and to a lesser extent, animal, human, and microorganism genomes (Table 14.1, Coassin et al. 2010; Garvin and Gharrett 2007; Till et al. 2006a; Yoshida et al. 2009). The potential for using Ecotilling for germplasm characterization and population genetics was first described in studies of the western black cottonwood *Populus trichocarpa* (Gilchrist et al. 2006). Taking advantage of recently available genome sequence and the fact that SNPs are a common form of diversity in most species, the authors evaluated polymorphisms in nine gene targets in a reference collection representing forty-one populations. To identify homozygous SNPs, an equal amount of reference DNA was added to samples before the Ecotilling reactions as was done for Arabidopsis. By comparing samples with and without reference DNA, the zygosity of SNPs were assigned, allowing the creation of a dendrogram of the sampled populations, something that cannot be done using heterozygous SNPs alone. An important limitation to the method of adding reference DNA was described in this work. Unlike the Columbia reference used for Arabidopsis Ecotilling, the *Populus* sample used as a reference harbours heterozygous SNPs. Homozygous SNPs in test samples cannot be inferred at positions where a heterozygous SNP occurs in the reference sample. This is because a band at that position is produced with or without the addition of the test sample. A strategy to overcome this limitation has been described for Ecotilling in the genus *Musa* (banana and plantains) where nascent heterozygosity is high (see below). Following what was done

in *Arabidopsis*, haplotypes of *Populus* were grouped by similar molecular weight, and only a subset sequence-validated. As expected, sequence validation revealed the majority of polymorphisms to be SNPs with two indels also recovered. From this, low nucleotide diversity was reported in comparison to other tree species, and it was estimated that populations were in Hardy-Weinberg equilibrium. Due to the visual nature of band analysis, unusual patterns were easily observed, and haplotype analysis revealed possible hybrid contaminants in the tested populations. This suggests Ecotilling could be used to quickly monitor the genotypic stability of germplasm in stock centers.

Following this work, a number of groups adapted Ecotilling for germplasm characterization in different plant species. As with black cottonwood, linkage disequilibrium in rice was estimated using Ecotilling (Rakshit et al. 2007). Barkley and colleagues developed an Ecotilling platform for characterization of *Vigna radiata* (mung bean) germplasm that is held by the USDA-ARS Plant Genetic Resources Conservation Unit (Barkley and Wang 2008). In this study, 157 polymorphisms including SNPs and indels representing 45 haplotypes were discovered in 22 tested samples. To maximize discoverable diversity, the authors took advantage of low selective pressure in noncoding regions of genes, and designed primers to amplify target sequences containing introns. Interestingly, the target amplified regions had a low mean GC content of 36%. By contrast, while some TILLING amplicons for maize had GC content of 70% (Till et al. 2007). This suggests that the standard polymorphism discovery platform of enzymatic mismatch cleavage followed by fluorescence detection is robust over a wide range of GC contents and thus applicable for most genomes and genomic regions.

Another potential bottleneck in establishing an Ecotilling platform is polyploidy. To avoid potential issues with co-amplification of related sequences, an approach using homeologue specific primers was developed for TILLING polyploid species such as wheat (Slade et al. 2005). Cooper and colleagues showed that the co-amplification of a duplicated region of the soybean genome resulted in increased false negative error rates in mutation discovery, and described a method of restriction digestion prior to target amplification to remove unwanted copies (Cooper et al. 2008). In polyploid potato, a homeologue-specific approach was taken that allowed for the development of markers to differentiate different Syrian potato cultivars (Elias et al. 2009). A similar tactic was taken in Brassica, where a homeologue-specific approach resulted in the development of genome-specific markers (Wang et al. 2010). Germplasm characterization was more challenging in the genus *Musa* where examples exist of diploid and triploid hybrids of the *acuminata* and *balbisiana* genome types. For *Musa* Ecotilling, the presence of diploid hybrids combined with high heterozygosity in cultivars limited the efficacy of a homeologue specific amplification approach. To overcome this limitation, primers were designed to specifically co-amplify target copies from both genomes with equal efficiency. To validate this approach, 48 accessions were screened in replicate using 7 gene targets, and 98% of bands were recovered in both replicates (Till et al. 2010). Sanger sequencing revealed a false negative error rate on Ecotilling gels of 6%, similar to that previously reported for

polymorphism recovery in human DNA samples (Till et al. 2006a). A total of 80 accessions from Bioversity's International Transit Centre and internal collections held by the IAEA Plant Breeding and Genetics Laboratory (PBGL) were then evaluated with 14 gene targets. Because high heterozygosity limited the use of a plant to serve as reference genotype, only heterozygous SNPs were recovered in the majority of the evaluations. Using a Principle Component Analysis, it was shown that evaluation of heterozygous SNPs alone was sufficient to differentiate hybrids from non-hybrids and triploid acuminata plants from diploids. Thus, a rapid SNP evaluation can in some cases replace flow cytometric methods used to differentiate ploidy. Further, differentiation between acuminata (AA) and balbisiana (BB) diploids was sufficient to uncover an accidental miss-assignment of an AA type as a BB type by the stock center. The authors further developed a strategy to exploit the high sensitivity of the standard fluorescence based Ecotilling assay by pooling representative genotypes so that a snapshot of all nucleotide diversity within a gene target could be evaluated in a single gel lane. To overcome the limitations of high heterozygosity in cultivars, a strategy utilizing doubled haploid plants as a reference is currently being developed (J. Jankowicz-Cieslak, B. Dussoruth, and B.J. Till, unpublished). Doubled haploid plants should be completely homozygous, thus providing the ideal reference sample. The efficacy of co-amplification of homeologous sequences for TILLING in triploid banana was also recently described, suggesting that this may be a suitable approach for induced mutation discovery in different polyploids (Jankowicz-Cieslak et al. 2012).

14.3.2 Ecotilling for Functional Genomics

A powerful approach to elucidating gene function is the discovery and evaluation of alleles in candidate genes predicted to control traits that vary between different genotypes. A variety of groups have used Ecotilling methods for this approach. In barley, allelic variation in genes known to be involved in resistance to fungal pathogens was investigated (Mejlhede et al. 2006). Mutant lines and cultivars were screened for SNPs and indels in the mildew resistant *mlo* gene and *Mla* locus. SNPs and indels were recovered and the authors proposed Ecotilling as an efficient genotyping platform for identifying and pyramiding useful alleles. Additional development of Ecotilling for plant disease resistance was performed in melon (Nieto et al. 2007). Nieto and colleagues evaluated allelic variation in eIF4E, a translation initiation factor previously implicated in response to virus infection in melon. They evaluated 113 accessions in which response to melon necrotic spot virus (MSNV) and Cucumber vein yellowing virus had been phenotypically evaluated. While most polymorphisms discovered were silent and are predicted to have no effect on protein function, one missense change was recovered that was found only in plants resistant to MNSV, suggesting a potential causal link. Following this work, screening eIF4E and eIFiso4E in 233 accessions of capsicum was performed to identify alleles conferring differential response to virus infection (Ibiza et al. 2010). Polymorphisms recovered were

validated in *in vitro* assays. A novel aspect of this work was the screening of cDNA sequences (described below). Taking a different approach to plant pathogen interactions, Yoshida and colleagues looked at nucleotide variation in the rice pathogen *Magnaporthe oryzae*, which is responsible for rice blast (Yoshida et al. 2009). A sample pooling strategy was employed and 46 isolates were screened in 1032 loci, with polymorphisms detected in 22 % of regions screened. Ecotilling was used in conjunction with whole genome sequencing and PCR based deletion screening to identify avirulence genes.

Abiotic stress, quality and other agronomic traits have also been investigated using Ecotilling. For example, genotypes conferring resistance to sulfonylurea herbicide resistance in *Monochoria vaginalis* were evaluated for polymorphisms in acetolactate synthase genes (Wang et al. 2007). In rice, the genetic component of salinity response has been studied. In this work, 28 SNPs and indels were recovered from two genes involved in salinity tolerance (Negrao et al. 2011). Increasing salinity is becoming an important issue in crop production in many countries (Pitman and Lauchli 2001). Efforts to identify genes controlling salinity resistance in rice have also been performed at the Plant Breeding and Genetics Laboratory of the FAO/IAEA Joint Programme. Here, 192 rice accessions were phenotypically evaluated for salt stress response using hydroponic glasshouse assays. This was followed by Ecotilling using target genes thought to play some role in salinity response. Over 60 sequence validated SNPs and indels were recovered and polymorphisms are being evaluated for correlations with stress response (S. Bado, O. Huynh and B.J. Till, unpublished). Like salinity tolerance, drought is a complex abiotic trait involving many genes. To investigate this, Yu and colleagues used Ecotilling to discover polymorphisms in the promoters of 24 transcription factor families to test for associations between nucleotide variation and drought tolerance (Yu et al. 2012). Alleles involved in protein quality in soybean were evaluated by using Ecotilling methods to screen five glycinin genes. The authors propose a combination approach of Ecotilling and Temperature-Switch PCR for marker assisted selection for breeding (Jegadeesan et al. 2012). In barley, 292 accessions were screened for variation of the *Lhcb1*/light harvesting chlorophyll A/B-binding protein, and 31 distinct haplotypes were recovered. Photosynthetic variation is potentially linked to agronomic traits such as yield. The authors therefore evaluated associations between *Lhcb1* polymorphisms and flag leaf area, leaf color, plant height, spike length, grains per spike and thousand grain weight (Xia et al. 2012).

While candidate gene approaches such as Ecotilling are most straightforward for monogenic traits, or traits under the control of only a few major genes, as described above, methods are also being employed for investigations into complex or quantitative trait loci (QTL). Ochai and colleagues used an innovative strategy to map and clone a QTL involved in resistance to increased levels of boron (Ochai et al. 2011). Toxicity to boron, typically introduced in water used for irrigation, causes decreased yield. Exploiting differential response to boron between indica and japonica rice cultivars, recombinant inbred lines were screened by Ecotilling and a QTL for boron toxicity was cloned. This was accomplished by first fine mapping boron toxicity response to approximately 49 kb. 44 japonica accessions were then screened

for boron response and grouped into sensitive and tolerant landraces. Ecotilling screens were performed spanning approximately 70 % of the 49 kb region. A total of eighty-seven primer pairs were designed to walk across the mapped region. The use of gene-specific fluorescently labelled primers that are common in Ecotilling assays would have been excessively costly. To avoid this, the authors used a low cost method employing unlabelled primers and agarose gels (see below). Several polymorphisms were recovered that were different between the two classes, most being in non-coding regions. However, a single nucleotide insertion was recovered in coding sequence. This insertion was also recovered in boron sensitive indica rice and predicted to cause a premature stop codon in a NAC like transcription factor. RNAi studies by the authors supported a role of this transcription factor in boron sensitivity.

14.4 Polymorphism Discovery Methods Adapted for Ecotilling

The development of low cost, accurate and high-throughput polymorphism discovery was a key feature enabling early success in Ecotilling. Unlike TILLING, which is a strategy that combines mutagenesis with mutation discovery, Ecotilling is often considered a methodology that uses enzymatic mismatch cleavage for the discovery and genotyping of natural nucleotide polymorphisms in populations. As such, TILLING has been described using denaturing High Pressure Liquid Chromatography (HPLC), enzymatic mismatch cleavage, High Resolution Melt (HRM) analysis, next generation sequencing, and other methods as the mutation discovery platform, while the majority of publications describing Ecotilling use enzymatic mismatch cleavage (Table 14.1, Jankowicz-Cieslak et al. 2011). Therefore, this section will focus only on adaptations to enzymatic mismatch cleavage as it relates to Ecotilling projects.

A large percentage of projects have used a crude celery juice extract (CJE) containing CEL I that was shown to have similar activity to purified CEL I and mung bean nucleases (Table 14.1). The popularity of this approach may in part be due to the relative ease of producing large amounts of enzyme from a small input of raw celery without the need of specialized equipment for biochemical purification. Additionally, high enzymatic cleavage activity has been shown over a wide range of incubation conditions such as salt concentrations, temperature and pH, making the cleavage assay robust and user friendly (accepting of fluctuations in assay parameters caused by human error) (Till et al. 2004a). Alternatives to CEL I or CJE that have been successfully used in TILLING or Ecotilling assays include commercially available Surveyor and ENDO I nucleases and crude enzyme extracts from brassica petioles (Nieto et al. 2007; Qiu et al. 2004; Sato et al. 2006). To date, evidence suggests that similar performance can be achieved from a variety of different enzymes and preparations, with variations in sensitivity and polymorphism discovery accuracy likely attributable to other assay parameters such as genomic DNA quality, primer design and sample pooling (Cooper et al. 2008; Till et al. 2004a; Till et al. 2004b).

An important caveat to the standard enzymatic mismatch cleavage assay is that it is biased for the discovery of SNPs and small indels. Deletions of 20–50 bp have been recovered (Comai et al. 2004, and B.J. Till, unpublished). While careful evaluation of the effect of indel size on mutation discovery has yet to be carried out, it is clear that at some point a deletion will not be detected because primer binding sites are also deleted, making PCR amplification impossible. For typical assays, amplicon sizes range between 750 bp to 1.5 kb, therefore indels larger than this size range will certainly go undetected. This may represent a minor ascertainment bias as Ecotilling projects are designed around evaluation of SNPs that represent the most common form of natural diversity. Therefore, for TILLING assays performed on material that harbours large deletions such as those induced by treatment with mutagens such as fast neutrons, alternative mutation discovery methods are required (Rogers et al. 2009).

Another area of protocol adaptation is in modification of the method for detecting cleaved DNA fragments. As with other aspects of Ecotilling, advances in this area have largely followed pioneering work in TILLING. The original platform used for Ecotilling, the LI-COR DNA analyser, provides a sensitive system that allows high-throughput polymorphism discovery in pooled samples, especially useful for the discovery of low frequency alleles (Till et al. 2006b). By using a self-prepared enzymatic extract for mismatch cleavage, the consumables costs are reduced, with the most expensive consumables being fluorescently end-labelled PCR primers. The cost of gene-specific primers can be approximately four times the cost of all other consumables combined when screening 96 or fewer samples. A large part of the primer cost is for synthesis, not molar primer amount, and so consumable costs drop appreciably with increasing sample size. Thus, the use of gene-specific labelled primers is economical for many TILLING projects where thousands of samples are routinely screened. As many Ecotilling projects involve the characterization of much fewer samples, gene-specific labelled primers can become cost prohibitive. To overcome this, universal labelling strategies have been developed whereby unlabelled gene-specific primers are designed to contain 5' sequences that are not complementary to the sampled genome. The sequences commonly used are the bacteriophage T3/7 or M13 promoter sequences (Rungis et al. 2005; Till et al. 2006a). IRDye labelled primers are then designed complementary to the universal sequences. These primers can then be used for all genes, dropping the price of labelled primer per 96 sample lanes from approximately 150 to 6 dollars. Such universal labelling has been used for polymorphism discovery projects in Brassica, rice, spruce and invasive weeds (Table 14.1). Universal priming strategies are broadly applicable to most fluorescent-based readout platforms including capillary electrophoresis (Rakshit et al. 2007).

An attractive alternative is to remove the need for fluorescent dyes. Perhaps the simplest permutation is the use of native agarose gel electrophoresis and DNA staining with dyes such as SYBR Green or ethidium bromide. Importantly, for such a method to work, both DNA strands must be cleaved at the site of the mismatch. Such an activity had been well established for the related single-strand specific mung bean nuclease (Kroeker et al. 1976). After digestion with nuclease, unpurified samples can be directly loaded directly onto agarose gels without further purification as is commonly used in the LI-COR based method. Mismatches in heteroduplexed molecules

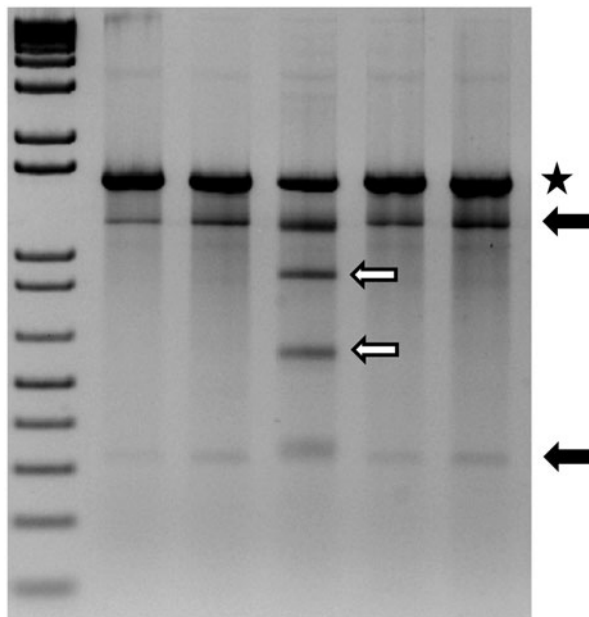


Fig. 14.2 Example of low cost Ecotilling in *Lupinus angustifolius*. PCR was performed on 5 accessions using primers amplifying a 1492 bp gene target. Samples were then subjected to cleavage with celery juice extract (CJE) and evaluated on a 1.5 % agarose gel stained with ethidium bromide. A cleavage event due to the presence of a SNP produces two bands whose molecular weights sum to the weight of the full-length PCR product (marked with a star). A common polymorphism is identified in all samples (resulting bands marked with black arrows). A rare polymorphism is found in one sample tested (marked with white arrows). This image was produced in the Plant Breeding and Genetics Laboratory (PBGL) of the IAEA Laboratories and kindly supplied by Joanna-Jankowicz-Cieslak of the PBGL and Kamila Kozak-Stankiewicz of the Department of Genetics, Plant Breeding and Seed Production, Wroclaw University of Environmental and Life Sciences, Poland

result in the appearance of two bands of lower molecular weight than the full length PCR product whose weights sum to that of the undigested product (Fig. 14.2). The primary disadvantages of this method are limited resolution and sensitivity as compared to fluorescent based methods. Sensitivity is most important when screening pooled samples, and so may be less important than limited resolution for most Ecotilling applications. Such limitations can lead to increased error rates. The precision of polymorphism discovery, however, may not be affected, thus making the agarose gel method suitable for rapid and low cost genotypic comparisons (Garvin and Gharrett 2007). A straightforward modification for increasing sample resolution can be achieved by replacing agarose with higher resolution polyacrylamide (Kadaru et al. 2006; Uauy et al. 2009). Non-electrophoretic readout platforms can also be considered in combination with enzymatic mismatch cleavage, such as denaturing HPLC, and HRM Analysis (Bono et al. 2011; Fusari et al. 2011; Gady et al. 2009).

While sample pooling is a major way to increase screening throughput, other methods can be considered. The size of PCR amplicons can be increased as has been described for rice Ecotilling where > 2 kb fragments were screened via agarose gels to map gene function (Raghavan et al. 2007). Presumably, amplicon size is limited only by the efficiency of PCR amplification and resolution of the readout platform. However, at some point increased thermal cycling and gel run times required for larger fragments may prove less efficient than screening a gene with multiple amplicons. Artifactual cleavage products due to mishybridization or template breathing within large duplexed molecules may also become an issue, as any single stranded region is a substrate for enzymatic cleavage. An alternative strategy to increase throughput is to limit polymorphism screening only to sequences of interest. For functional genomics applications where nucleotide polymorphisms in the coding sequence are primarily sought, screening cDNAs can dramatically increase efficiency, especially in species where genes contain frequent or large intronic sequences. For example, cDNA sequences in 4 gene targets were screened by Ecotilling to evaluate polymorphisms associated with virus resistance in peppers (Ibiza et al. 2010). While this required the additional steps of first strand synthesis and product purification, screening efficiency was effectively doubled as evaluation of intronic sequence was avoided. The approach may not be suitable for diversity studies where maximizing discovery of polymorphic sites is desired, due to the fact that the frequency of SNPs and indels is typically higher in non-coding regions where there is lower selective pressure. An interesting modification to this approach would be the generation of total cDNA libraries to enable rapid screening of all transcribed genes. This had been discussed for many years in the TILLING community, but to the knowledge of the author, never published. Issues with assay consistency from transcripts of varying abundance, and also nonsense mediated mRNA decay that could remove desired induced knockouts could potentially limit the efficacy of this approach.

14.5 Future Directions

When considering the future of nucleotide polymorphism studies in plants, it is clear that novel mutation technologies that provide higher throughput at a lower cost will eventually replace enzymatic mismatch cleavage based approaches. This can already be observed in the work of large labs or consortia where resources exist to cover the high initial cost of new equipment and the development of appropriate informatics infrastructure. The dominant technologies in the near future are likely to be one of several next or new generation sequencing platforms (see the chapter by David Edwards, this book). For example, sequencing based efforts are ongoing to catalogue diversity in 1000 human genomes, and 1001 Arabidopsis accessions (Altshuler et al. 2010; Weigel and Mott 2009). Discoveries from these projects such as the nature of gene by environment interactions, evolutionary pressures, and the establishment of robust technology platforms will no doubt lead to similar efforts in crops. However, prior to the availability of truly low cost whole genome sequencing for large plant

genomes, the value of whole genome versus targeted amplicon approaches must be weighed against available resources and research objectives. In many cases, amplicon selection, multiplexing and sample pooling can provide a robust approach for polymorphism discovery (Tsai et al. 2011). The use of this approach for natural polymorphisms has been termed next-generation Ecotilling (Marroni et al. 2011). While the pre selection or filtration of genomic sequences prior to sequencing reduces costs and informatics load, there will likely come a time when whole genome sequencing costs and informatics tools are at a point where whole genome evaluation is common and available to most labs. Until that day comes, lower throughput and lower cost applications such as traditional Ecotilling will remain useful tools for smaller scale projects where resources are limited, or outsourcing is impractical. Ultra low cost methods such as agarose gel Ecotilling will likely remain important tools in developing countries that are rich in largely uncharacterized local germplasm.

Acknowledgements Work described in this review on Musa, Lupin and rice Ecotilling performed in the Plant Breeding and Genetics Laboratory was funded by the Food and Agriculture Organization of the United Nations and the International Atomic Energy Agency through their Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture. The author thanks Rachel Howard-Till and Brian Forster for careful review of the manuscript and providing useful suggestions for its improvement. The author also wishes to thank members of the TILLING and Ecotilling community for stimulating conversations over the years and regrets any accidental omissions of Ecotilling work from this review.

References

- Altshuler DL, Durbin RM, Abecasis GR et al (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073
- Barkley NA, Wang ML (2008) Application of TILLING and EcoTILLING as reverse genetic approaches to elucidate the function of genes in plants and animals. *Curr Genomics* 9:212–226
- Bono C, Nuzzo D, Albegiani G et al (2011) Genetic screening of Fabry patients with EcoTILLING and HRM technology. *BMC Res Notes* 4:323
- Caicedo AL, Purugganan MD (2005) Comparative plant genomics. *Frontiers and prospects. Plant Physiol* 138:545–547
- Chen LA, Wang SQ, Hu YG (2011) Detection of SNPs in the VRN-A1 gene of common wheat (*Triticum aestivum* L.) by a modified Ecotilling method using agarose gel electrophoresis. *Aust J Crop Sci* 5:318–326
- Coassin S, Schweiger M, Kloss-Brandstatter A et al (2010) Investigation and functional characterization of rare genetic variants in the adipose triglyceride lipase in a large healthy working population. *Plos Genet* 6
- Colbert T, Till BJ, Tompa R et al (2001) High-throughput screening for induced point mutations. *Plant Physiol* 126:480–484
- Comai L, Young K, Till BJ et al (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *Plant J* 37:778–786
- Cooper JL, Till BJ, Laport RG et al (2008) TILLING to detect induced mutations in soybean. *BMC Plant Biol* 8:9
- Cordeiro G, Elliott FG, Henry RJ (2006) An optimized ecotilling protocol for polyploids or pooled samples using a capillary electrophoresis system. *Anal Biochem* 355:145–147

- Elias R, Till BJ, Mba C, Al-Safadi B (2009) Optimizing TILLING and Ecotilling techniques for potato (*Solanum tuberosum* L). *BMC Res Notes* 2:141
- Fusari CM, Lia VV, Nishinakamasu V et al (2011) Single nucleotide polymorphism genotyping by heteroduplex analysis in sunflower (*Helianthus annuus* L.). *Mol Breeding* 28:73–89
- Gady AL, Hermans FW, Van de Wal MH et al (2009) Implementation of two high through-put techniques in a novel application: detecting point mutations in large EMS mutated plant populations. *Plant Methods* 5:13
- Galeano CH, Gomez M, Rodriguez LM, Blair MW (2009) CEL I Nuclease Digestion for SNP Discovery and Marker Development in Common Bean (*Phaseolus vulgaris* L.). *Crop Sci* 49:381–394
- Garvin MR, Gharrett AJ (2007) DEco-TILLING: an inexpensive method for single nucleotide polymorphism discovery that reduces ascertainment bias. *Mol Ecol Notes* 7:735–746
- Gilchrist EJ, Haughn GW, Ying CC et al (2006) Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Mol Ecol* 15:1367–1378
- Greene EA, Codomo CA, Taylor NE et al (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. *Genetics* 164:731–740
- Ibiza VP, Canizares J, Nuez F (2010) EcoTILLING in *Capsicum* species: searching for new virus resistances. *Bmc Genomics* 11
- Jankowicz-Cieslak J, Huynh OA, Bado S et al (2011) Reverse-genetics by TILLING expands through the plant kingdom. *Emirates J Food Agric* 23:290–300
- Jankowicz-Cieslak J, Huynh OA, Brozyska M et al (2012) Induction, rapid fixation and retention of mutations in vegetatively propagated banana. *Plant Biotechnol J* doi: 10.1111/j.1467-7652.2012.00733.x. [Epub ahead of print]
- Jegadeesan S, Yu K, Woodrow L et al (2012) Molecular analysis of glycinin genes in soybean mutants for development of gene-specific markers. *Theor Appl Genet* 124:365–372
- Kadar SB, Yadav AS, Fjellstrom RG, Oard JH (2006) Alternative ecotilling protocol for rapid, cost-effective single-nucleotide polymorphism discovery and genotyping in rice (*Oryza sativa* L.). *Plant Mol Biol Rep* 24:3–22
- Kaul S, Koo HL, Jenkins J et al (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Kroeker WD, Kowalski D, Laskowski M Sr (1976) Mung bean nuclease I. Terminally directed hydrolysis of native DNA. *BioChemistry* 15:4463–4467
- Marroni F, Pinosio S, Di CE (2011) Large-scale detection of rare variants via pooled multiplexed next-generation sequencing: towards next-generation Ecotilling. *Plant J* 67:736–745
- McCallum CM, Comai L, Greene EA, Henikoff S (2000a) Choosing optimal regions for TILLING. *Plant Physiol* 123:439–442
- McCallum CM, Comai L, Greene EA, Henikoff S (2000b) Targeted screening for induced mutations. *Nat Biotechnol* 18:455–457
- Mejlhede N, Kyjovska Z, Backes G et al (2006) EcoTILLING for the identification of allelic variation in the powdery mildew resistance genes *mlo* and *Mla* of barley. *Plant Breeding* 125:461–467
- Negrao S, Almadanim C, Pires I et al (2011) Use of EcoTILLING to identify natural allelic variants of rice candidate genes involved in salinity tolerance. *Plant Genet Resour-C* 9:300–304
- Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814
- Nieto C, Piron F, Dalmais M et al (2007) EcoTILLING for the identification of allelic variants of melon eIF4E, a factor that controls virus susceptibility. *BMC Plant Biol* 7:34
- Ochiai K, Shimizu A, Okumoto Y et al (2011) Suppression of a NAC-Like Transcription Factor Gene Improves Boron-Toxicity Tolerance in Rice. *Plant Physiol* 156:1457–1463
- Oleykowski CA, Bronson Mullins CR, Godwin AK, Yeung AT (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Res* 26:4597–4602

- Pitman MG, Lauchli A (2001) Global Impact of Salinity and Agricultural Ecosystems. In: Lauchli A, Lutttge U (eds) *Salinity: Environment-Plants-Molecules*. Kluwer Academic Publishers, pp 3–20
- Qiu P, Shandilya H, D’Alessio JM et al (2004) Mutation detection using Surveyor nuclease. *Biotechniques* 36:702–707
- Raghavan C, Naredo MEB, Wang HH et al (2007) Rapid method for detecting SNPs on agarose gels and its application in candidate gene mapping. *Mol Breeding* 19:87–101
- Rakshit S, Rakshit A, Matsumura H et al (2007) Large-scale DNA polymorphism study of *Oryza sativa* and *O. rufipogon* reveals the origin and divergence of Asian rice. *Theor Appl Genet* 114:731–743
- Ramos ML, Huntley JJ, Maleki SJ, Ozias-Akins P (2009) Identification and characterization of a hypoallergenic ortholog of *Ara h 2.01*. *Plant Mol Biol* 69:325–335
- Rogers C, Wen J, Chen R, Oldroyd G (2009) Deletion based reverse genetics in *Medicago truncatula*. *Plant Physiol* 151:1077–1086
- Rungis D, Hamberger B, Berube Y et al (2005) Efficient genetic mapping of single nucleotide polymorphisms based upon DNA mismatch digestion. *Mol Breeding* 16:261–270
- Sato Y, Shirasawa K, Takahashi Y et al (2006) Mutant Selection from Progeny of Gamma-ray-irradiated Rice by DNA Heteroduplex Cleavage using Brassica Petiole Extract. *Breeding Science* 56:179–183
- Slade AJ, Fuerstenberg SI, Loeffler D et al (2005) A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Nat Biotechnol* 23:75–81
- Taylor NE, Greene EA (2003) PARSESNP: A tool for the analysis of nucleotide polymorphisms. *Nucleic Acids Res* 31:3808–3811
- Till BJ, Burtner C, Comai L, Henikoff S (2004a) Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Res* 32:2632–2641
- Till BJ, Reynolds SH, Weil C et al (2004b) Discovery of induced point mutations in maize genes by TILLING. *BMC Plant Biol* 4:12
- Till BJ, Zerr T, Bowers E et al (2006a) High-throughput discovery of rare human nucleotide polymorphisms by Ecotilling. *Nucleic Acids Res* 34:e99
- Till BJ, Zerr T, Comai L, Henikoff S (2006b) A protocol for TILLING and Ecotilling in plants and animals. *Nat Protoc* 1:2465–2477
- Till BJ, Cooper J, Tai TH et al (2007) Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biol* 7:19
- Till BJ, Jankowicz-Cieslak J, Sagi L et al (2010) Discovery of nucleotide polymorphisms in the *Musa* gene pool by Ecotilling. *Theor Appl Genet* 121:1381–1389
- Tsai H, Howell T, Nitcher R et al (2011) Discovery of Rare Mutations in Populations: TILLING by Sequencing. *Plant Physiol* 156:1257–1268
- Uauy C, Paraiso F, Colasuonno P et al (2009) A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. *BMC Plant Biol* 9:115
- Wang GX, Tan MK, Suj ARC et al (2007) Discovery of single-nucleotide mutations in acetolactate synthase genes by Ecotilling. *Pestic Biochem Phys* 88:143–148
- Wang NA, Shi L, Tian F et al (2010) Assessment of *FAEI* polymorphisms in three Brassica species using EcoTILLING and their association with differences in seed erucic acid contents. *Bmc Plant Biol* 10
- Weigel D, Mott R (2009) The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biol* 10
- Xia Y, Ning Z, Bai G et al (2012) Allelic variations of a light harvesting chlorophyll a/b-binding protein gene (*Lhcb1*) associated with agronomic traits in barley. *PLoS One* 7:e37573
- Yoshida K, Saitoh H, Fujisawa S et al (2009) Association Genetics Reveals Three Novel Avirulence Genes from the Rice Blast Fungal Pathogen *Magnaporthe oryzae*. *Plant Cell* 21:1573–1591
- Yu S, Liao F, Wang F et al (2012) Identification of rice transcription factors associated with drought tolerance using the Ecotilling method. *PLoS One* 7:e30765
- Zerr T, Henikoff S (2005) Automated band mapping in electrophoretic gel images using background information. *Nucleic Acids Res* 33:2806–2812

Part III
Genome Sequencing and Crop
Domestication

Chapter 15

Next Generation Sequencing and Germplasm Resources

Paul Visendi, Jacqueline Batley and David Edwards

Contents

15.1 Introduction	370
15.2 Next Generation Sequencing	372
15.3 Sequencing Reference Genomes	373
15.4 Next Generation Diversity Analysis	379
15.4.1 Reference-Based Diversity Analysis	381
15.4.2 Non Reference-Based Diversity Analysis	381
15.5 SNPs	381
15.5.1 Copy Number Variation	383
15.6 Perspectives	384
References	384

Abstract DNA sequencing technology is advancing at an astounding rate, with rapid increases in data volumes and quality combined with reducing costs. The availability of this technology opens novel avenues for the analysis of plant germplasm resources. Where previous studies analysed a limited number of phenotypic or molecular genetic markers, it is now possible to re-sequence whole genomes to characterise diversity at a resolution of each nucleotide. Current approaches combine high resolution genetic markers with genome sequencing both for reference assembly and genotyping by sequencing. As next generation sequencing technologies continue to advance, we approach the potential to catalogue and characterise all genome variations across

D. Edwards (✉) · P. Visendi · J. Batley
University of Queensland, School of Agriculture and Food Sciences,
4072 Brisbane, QLD, Australia
e-mail: Dave.Edwards@uq.edu.au

D. Edwards · P. Visendi
Australian Centre for Plant Functional Genomics, University of Queensland,
4072 Brisbane, QLD, Australia

P. Visendi
e-mail: paul.muhindira@uqconnect.edu.au

J. Batley
e-mail: J.Batley@uq.edu.au

diverse germplasm to gain a greater understanding of how the genome contributes to the diversity seen in today's plants.

Keywords Genome sequencing · Pangenome · Illumina · Ion Torrent · AB SOLiD · Pacific biosciences · Oxford nanopore · Roche 454 · Single nucleotide polymorphisms · Bioinformatics

List of species

- *Arabidopsis thaliana*
- *Brachypodium distachyon*
- *Brassica juncea*
- *Brassica oleracea*
- *Brassica rapa*
- *Carica papaya*
- *Eucalyptus grandis*
- *Fragaria vesca*
- *Glycine max*
- *Lotus japonicus*
- *Malus domestica*
- *Medicago truncatula*
- *Mimulus guttatus*
- *Oryza sativa*
- *Populus trichocarpa*
- *Prunus persica*
- *Solanum tuberosum*
- *Sorghum bicolor*
- *Theobroma cacao*
- *Triticum aestivum*
- *Vitis vinifera*
- *Zea mays*

15.1 Introduction

Plant genomes are often both highly repetitive with ancient and more recent polyploidy. Because of these features, taxonomic analysis can be a challenge. To comprehensively analyse germplasm resources, allele and haplotype frequencies need to be studied. Next Generation Sequencing (NGS) techniques have greatly advanced genome analysis, enabling the elucidation of the genome or transcriptome sequence of an organism relatively quickly and cheaper in comparison to the traditional Sanger-based sequencing technologies (Edwards et al. 2013). This has opened avenues in plant research, by which genomic variations can be quickly and efficiently determined through computational and statistical analysis.

While NGS data is relatively inexpensive compared to traditional sequence data, sequence quality and read lengths are relatively poor. Accurate estimation of Single Nucleotide Polymorphism (SNP) calling and allele frequencies may result in false positives due to sequence or read mapping error. In some cases, a greater number of biological replicates rather than depth of coverage has been shown to produce better results when calling SNP and allele frequencies (Li 2011). These errors require downstream validation and confirmatory analysis, which usually involves re-sequencing (Kim et al. 2011).

NGS technologies have impacted on plant taxonomic systems, currently leveraging the Linnaean taxonomic methods with DNA barcoding (Pang et al. 2012) and evolutionary studies in plants (Darracq et al. 2010). Methods which make this possible for complex polyploid genomes include targeted sequence capture (Grover et al. 2012) and chromosome flow cytometry (Doležel et al. 2004). These approaches reduce the complexity of the assembly of NGS data as well as downstream analysis, enabling biologists to answer questions on diversity, polyploid origins, domestication and ancestry of many crops (Zhang et al. 2011).

With the development of advanced sequencing systems, sequence-based SNPs are becoming the marker of choice. The greatest genetic resolution is obtained through the analysis of SNPs and allele frequencies within a population. Two alleles associating with each other by random chance are said to be in Linkage Equilibrium (LE). If this association is found to be non-random either due to physical proximity or selection, they are said to be in Linkage Disequilibrium (LD). LD is the basis of Association Mapping (AM). Accurate AM is highly dependent on the extent of physical LD, population size and population structure (Duran et al. 2010). Small and highly structured populations often lead to elevated false positives in AM studies (Cuesta-Marcos et al. 2010).

The dynamic nature of living organisms, their interaction with the environment, with other species and with each other, affects adaptation and evolution. This led to the “Pangenome” concept (Tetz 2005). This concept proposes that genetic information of all living organisms belongs to a common system, the Pangenome, and that this system is both stable and fluid, and genetic elements such as DNA, RNA and plasmids among others traverse through the system, implying horizontal gene transfer between similar organisms and even across species. Gene transfer has been reported between bacteria (Yue et al. 2012), between viruses and eukaryotes (Wu and Zhang 2011), between plants and prokaryotes, and plants and other eukaryotes (Bock 2010). The concept further postulates that mechanisms of fluidity of the Pangenome include viruses and bacteria through infection, the food chain, death and decay. There has been little advancement of this concept in plants but several studies have been carried out in humans looking at the Asian, Caucasian and African human genomes (Li et al. 2010a) and in bacteria (Hall et al. 2010; Laing et al. 2010). It would be interesting to see applications of this concept to higher plants.

15.2 Next Generation Sequencing

DNA sequencing technology is undergoing a revolution and at the same time fuelling a revolution in genetics and genomics. Applications for Sanger-based sequencing remain, though the majority of DNA sequencing is now produced by one of a range of NGS technologies. NGS suffers from shorter reads and greater error rates than traditional Sanger sequencing, and its predominance is due to the ability to produce much larger volumes of data at a relatively low cost per sequenced base.

The first commercially available pyrosequencing system was commercialised by Roche (Basel, Switzerland) as the GS20, capable of sequencing over 20 million base pairs (Mbp) in just over 4 h (Margulies et al. 2005). This was replaced in 2007 by the GS FLX model, capable of producing over 100 Mbp of sequence in a similar amount of time. The current system, the GS FLX + can produce around 700 Mbp of data with read lengths of up to 1,000 bp with multiplexing of samples (www.my454.com). The Roche 454 FLX system performs amplification and sequencing in a high-throughput picoliter format. Emulsion PCR enables the amplification of a DNA fragment immobilized on a bead, generating sufficient DNA for the subsequent sequencing reaction. Beads are distributed onto the plate. DNA sequencing involves the sequential flow of both nucleotides and enzymes over the plate, which converts chemicals generated during nucleotide incorporation into a chemiluminescent signal that can be detected by a CCD camera. The light signal is quantified to determine the number of nucleotides incorporated during the extension of the DNA sequence. The output is in the form of 'flow space', which is converted to the traditional ACGT nucleotide sequence format. The sequence reads are much longer than most other NGS systems. The main error types are additional or reduced numbers of nucleotides around mononucleotide strings. These errors make the accurate calling of insertion/deletion (indel) differences a challenge.

The Illumina sequencing platforms use reversible terminator chemistry to generate up to 600 Gbp of sequence data per run, the greatest volume of data from any current NGS platform. Sequencing templates are immobilized on a flow-cell surface, and amplification generates clusters of up to 1,000 identical copies of each DNA molecule. Sequencing uses fluorescently labelled nucleotides to produce reads of up to 150 bp in length, though 100 bp is more common. Reads can be produced as pairs. The use of paired reads improves the accuracy of reference mapping, overcoming many of the limitations of short read lengths such as inaccurate resolution of repeats, indels and structural rearrangements. By using the distance between a read pair to infer an insertion or deletion in the reference or sample and to resolve repeats in *de novo* assembly, higher accuracy is achieved. Illumina sequencing is now becoming the platform of choice for resequencing, SNP discovery, whole-genome shotgun sequencing and *de novo* assembly (Imelfort et al. 2009b; Imelfort and Edwards 2009; Williams-Carrier et al. 2010; Dong et al. 2011; Shulaev et al. 2011).

The SOLiD System from Life Technologies (Applied Biosystems) enables parallel sequencing of amplified DNA fragments linked to beads. The method uses

sequential ligation of dye-labelled oligonucleotides, and the latest 5500xl system produces 20–30 Gbp of data per day, with read lengths of up to 75 bp (www.appliedbiosystems.com/). SOLiD data features a two-base encoding mechanism that interrogates each base twice providing a form of built-in error detection for SNP discovery when comparing reads to a reference.

Ion Torrent is a relatively new technology and uses a high-density array of semiconductor micro reaction chambers (www.iontorrent.com). Changes in pH are recorded as a result of the release of a hydrogen proton during the incorporation of a nucleotide during DNA synthesis. This produces reads of 100–200 bp, with up to 1 Gbp of data per run. The error profile of this system is still unknown, but the technology has potential for cost-effective re-sequencing and variant discovery with fast runs of 2 h.

Pacific Biosciences is one of the first ‘third generation’ sequencing systems to go on the market, and applies a novel single-molecule sequencing technique called SMRT™ (Single Molecule Real Time) technology. Read lengths of around 1,000 bp have been reported (www.pacificbiosciences.com). As with the Ion Torrent system, little is known about the error profile of the system, but missing bases, and hence indel calling would be a likely challenge with this technology.

NGS technologies continue to evolve at an astounding rate and new technologies such as the Oxford Nanopore system are likely to continue to push the market forward over the coming years.

The vast quantities of data generated using NGS require the development of dedicated bioinformatics systems (Edwards et al. 2009; Marshall et al. 2010; Lai et al. 2012c; Lee et al. 2012). It was initially thought that bioinformatics systems would not be able to keep pace with sequencing developments, and while still a bottleneck exists in translating the data into biological information, the growth of bioinformatics has kept track with data production (Batley and Edwards 2009a; Lee et al. 2011b).

15.3 Sequencing Reference Genomes

There are now several optional approaches to sequence genomes, and the approach undertaken would depend on the use of the genome sequence (Imelfort et al. 2009a; Edwards and Batley 2010). Bacterial Artificial Chromosome (BAC) sequencing is still considered to be the most robust method for genome sequencing and involves the production of an overlapping tiling path of large genomic fragments maintained within BACs. Each BAC is shotgun sequenced, where many short reads are assembled to produce the sequence of the BAC. The whole genome sequence may then be reassembled based on sequence overlaps. This approach, while being the gold standard, remains prohibitively expensive and is unlikely to be undertaken for the majority of germplasm resources.

An alternative approach is the whole-genome shotgun (WGS) method, where the genome is fragmented into millions of smaller reads that are individually sequenced.

Computational algorithms then assemble the genome sequence, frequently requiring additional scaffolding or assembly to generate a representative genome. Scaffolding involves the assembly of several overlapping contigs into scaffolds and then assembling these further into a genome assembly. While WGS requires less time and resources than a BAC-by-BAC approach, assembly is often problematic due to repeats within the genome. This is particularly true for many plant species, with polyploid genomes and abundant repetitive elements (e.g. maize, wheat, etc.). With the decreasing cost of sequencing and rapid improvements in both data quality and read length from next and third generation technologies, WGS sequencing projects are likely to be established for many additional germplasm resources.

Arabidopsis thaliana was the first plant species to have a sequenced genome (Arabidopsis_Genome_Initiative 2000). Since this milestone, the number of plant genomes being sequenced continues to increase. While several species with small genome sizes as well as model species are now been sequenced, genome researchers are now starting to tackle some of the larger and more complex genomes (Berkman et al. 2011b; Feuillet et al. 2011; Edwards and Wang 2012; Berkman et al. 2013). These can act both as models to understand broad families of plants, as well as reference sequences for the mapping and comparison of unassembled genome sequence data from diverse germplasm resources.

Brassica species share extensive synteny with *Arabidopsis thaliana*, enabling comparative mapping and exploitation of the Arabidopsis genome sequence for Brassica crop improvement. Among the six cultivated species of Brassica, *B. rapa* (syn. *campestris*, AA, $n = 10$), *B. juncea* (AABB, $n = 18$) and *B. napus* (AACC, $n = 19$) are agronomically important oilseed crops, whereas *B. oleracea* (CC, $n = 9$) provides valuable leafy vegetables (e.g. broccoli, cauliflower, cabbage, kholrabi, etc.). The other two species, *B. nigra* (BB, $n = 8$) and *B. carinata* (BBCC, $n = 17$) are largely valued as condiments. Proprietary AA, CC and AACC genomes were sequenced in 2009 (<http://tinyurl.com/brassicagenome>), and recently the multinational Brassica genome project (MBGP) published the first public *B. rapa* genome (Mun et al. 2010; Wang et al. 2011; Edwards and Wang 2012).

Legumes represent the third largest plant family and are the second most important crop family for the human diet (Cannon et al. 2006). The model species *Medicago truncatula* is an annual diploid with eight chromosomes. It is closely related to tetraploid alfalfa (*M. sativa*). A combination of cytogenetic and BAC sequence data show that the *M. truncatula* genome is organized into distinct gene-rich euchromatin and repeat-rich pericentromeric regions, allowing the *M. truncatula* genespace to be efficiently sequenced using a BAC-by-BAC strategy (Young et al. 2011). Six chromosomes have been sequenced in a US project and two additional ones were sequenced by partners in Europe. Of special note is the *Medicago* HapMap Project, which aims to deep-sequence whole genomes of 30 inbred *Medicago* lines using the Illumina platform, and use the reference genome to determine SNPs and Indels. A current release (Mt3.5) of the genome is available at (www.medicago-hapmap.org).

Lotus japonicus is a diploid self-fertile perennial pasture legume, with six chromosomes and a genome sequence of around 450 Mbp. Large-scale genome sequencing of variety Miyakojima MG-20 began in 2000. ESTs, cDNAs and gene segments

from *Lotus* and other legumes were used to select TAC clones, which were then sequenced using a shotgun approach (Sato et al. 2008). Data released is available at www.kazusa.or.jp/lotus. The genome sequences of *Medicago truncatula* and *Lotus japonicus* have provided invaluable resources for legume research given that they have both been sequenced using BAC clones making syntenic studies relatively simple and in turn providing evolutionary insights into other species such as *Arabidopsis* (George et al. 2008; Schlueter et al. 2008; Bertioli et al. 2009).

Glycine max (Soybean) is a major crop that accounts for 70 % of the world's edible plant-derived protein. Its 1.1 Gbp genome was sequenced using a WGS approach (Schmutz et al. 2010a), and the sequence is available through the phytozome database (<http://www.phytozome.net/soybean>). Re-sequencing of cultivated and wild varieties has enabled a detailed characterisation of genome variation in this species (Lam et al. 2010). Early and current studies on the evolution and domestication of soybean have shown a loss of genetic diversity as a result of domestication (Hyten et al. 2006; Li et al. 2010b). It has also been hypothesized that there was a single soybean domestication event, although this has not been confirmed since archaeological evidence suggests multiple domestication sites in East Asia (Lee et al. 2011a).

Trees, due to their long life span, have characteristics that distinguish them from annual, herbaceous plants. It is likely that many of these properties are based on a tree-specific genetic foundation. Poplar has been selected as a model system for trees because it has a relatively small genome. *Populus trichocarpa* (black cottonwood), with a paleopolyploid ($2n = 38$) genome of approximately 480 Mbp, was selected as the first tree genome to be sequenced. The *Populus trichocarpa* Nisqually 1 genotype was sequenced using WGS approaches (Tuskan et al. 2006). Approximately 7.6 million reads were assembled into 2,447 scaffolds containing 410 Mb of genomic DNA. The assembly was found to include more than 95 % of known cDNAs. Sequencing this genome allowed for the comparison between perennial and annual plant species on a whole genome basis for the first time and provides resources to help answer tree-specific questions about dormancy, development of a secondary cambium, juvenile-mature phase change and long-term host-pest interactions (Brunner et al. 2004; Tuskan et al. 2004).

Among the hardwoods, the most widely grown is *Eucalyptus*. *Eucalyptus* has high economic value due to its fiber, largely exploited for pulp, cellulose and paper. In addition, *Eucalyptus* has potential as a sustainable biofuel source, due to a fast growth rate. In June 2007 the DOE JGI initiated the *Eucalyptus grandis* genome-sequencing project. The project started in August 2007, and sequencing was completed by 2009. The project, coordinated by the *Eucalyptus* Genome Network (EUCAGEN), involved more than 130 scientists from 18 countries. A release of the first annotated assembly is available on phytozome (<http://www.phytozome.net/>), with an approximate genome size of 691 Mb, 4,952 scaffolds from 32,762 contigs and 300 scaffolds greater than 50 kb in size, representing approximately 94.2 % of the genome.

In 2004, the tomato genome was selected as the reference for sequencing within the Solanaceae genomics project. The sequencing of the domesticated tomato (*Solanum lycopersicum*) (Consortium 2012) marks the first step in bringing together genetic maps and genomes of all Solanaceae and related plants, including potato, eggplant,

pepper, petunia and coffee. The variety 'Heinz 1706' was initially selected as BAC resources were already available. The 900 Mbp genome was sequenced using a combination of sanger and WGS approach. While the genome was known to consist of approximately three-quarters pericentromeric heterochromatin, known to be rich in repetitive sequences and poor in genes, the remaining one-quarter consisted of distal, euchromatic segments of chromosomes that contain mostly single copy sequences and more than 90 % of the genes. As such, only this portion was sequenced (Shibata 2005). To gain further insight into its evolution, its closest wild relative, *Solanum pimpinellifolium* LA1589 was also sequenced using illumina short reads (Tomato Genome Consortium 2012). The resulting 739 Mb draft genome showed a 0.6 % divergence from the domesticated variety. Comparative studies with potato (Xu et al. 2011c) revealed an 8.7 % nucleotide divergence. Further comparisons with the grape genome supported previous hypothesis that the rosoid lineage diverged from a common eudicot ancestor following whole genome triplication. The tomato genome was however shown to have high synteny with potato, pepper, eggplant and nicotiana (Tomato Genome Consortium 2012).

Potato, another member of the family Solanaceae has a variety of ploidy levels, ranging from diploid ($2n = 24$) to hexaploid ($6n = 72$), with the cultivated potato varieties being tetraploid. Potato is an economically important food crop with 330 million tons produced globally in 2009 (<http://www.fao.org>). The size of the genome is 840 Mbp. The potato genome sequencing consortium (PGSC) sequenced the potato genome using a BAC-by-BAC approach (Xu et al. 2011b). Due to its high heterozygosity, the generation of a draft sequence required the use of a homozygous form, called a doubled monoploid (DM), and integration of the sequence with that of a heterozygous diploid form RH89-039-16. The two genomes were used to study the genome structure, with the heterozygous diploid resembling the cultivated tetraploid potato (http://solgenomics.net/organism/Solanum_tuberosum/genome).

The sequencing of grapevine *Vitis vinifera* ($2n = 38$) (Jaillon et al. 2007), was performed on the quasi-homozygous genotype PN40024 by a French-Italian collaborative project using a WGS approach. A total of 6.2 million reads representing 8.4x coverage of the genome were assembled with the Arachne12 assembler to produce 316 supercontigs, representing putative allelic haplotypes totaling 11.6 Mbp. The assembly of one of the haplotypes in each heterozygous region resulted in 19,577 contigs and 3,514 supercontigs totaling 487 Mb, consistent with an earlier predicted size of 475 Mb (Lodhi et al. 1995). A different approach to sequencing *V. vinifera* used a BAC-by-BAC approach (Zharkikh et al. 2008). In a comparison of the two approaches (Zharkikh et al. 2008) showed it was possible to sequence the highly heterozygous genome by using sufficient coverage and a variety of clones with different sizes. This approach yielded more informative data on repetitive elements and useful SNPs.

The monkey flower *Mimulus guttatus*, has become a model system for studying ecological and evolutionary genetics due to its diverse phenotypes, which include adaptations to desert and aquatic environments, selfing and outcrossing, annual and perennial forms and varied floral morphology. The DOE Joint Genomes Institute (JGI) commenced sequencing *Mimulus guttatus* in 2006 using a WGS approach.

In addition to the WGS sequence, JGI is sequencing 200,000 ESTs each from *M. guttatus* and *M. lewisii*. Additionally, WGS sequencing of IM62 inbred lines is ongoing. A draft release of the genome with 321.7 Mb of 2,216 scaffolds, 300.7 Mb of 17,831 contigs with gaps $\sim 6.5\%$ and 512 scaffolds larger than 50 Kb, with 95.7% of the genome represented in scaffolds greater than 50 Kbp is available on phytozome (<http://www.phytozome.net/mimulus>).

The papaya-sequencing project was founded by the centre for genomics, proteomics and bioinformatics research Initiative (CGPBRI) at the University of Hawaii in 2004. Papaya (*Carica papaya*) is the first fruit species and commercially important transgenic plant to be sequenced (Ming et al. 2008). Papaya has nine chromosomes and the size of the genome is 372 Mbp. A WGS approach produced a total of 2.8 million reads generated from a female transgenic cultivar “SunUp”. After filtering, 1.6 million reads were assembled into contigs containing 271 Mb and scaffolds spanning 370 Mb. Validation of the assembly using 16,362 unigenes derived from expressed sequence tags (ESTs), showed 15,064 ESTs (92.1%) matched the assembly.

Cocoa (*Theobroma cacao*; $2n = 2x = 20$) was sequenced by the international cocoa genome sequencing consortium (ICGS), using a WGS approach with Sanger and Roche 454 technologies (Argout et al. 2011). Assembly was carried out using Newbler software, resulting in 25,912 contigs and 4,792 scaffolds, with an N50 of 473.8 Kb. Illumina reads (x44 coverage) were used to improve the assembly. The total assembly length was 326 Mb, representing approximately 76% of the estimated genome size of cocoa. The sequencing of cocoa enabled the comparison of the grape, soybean, poplar and *A. thaliana* genomes with cocoa revealing 682 gene families (2,053 genes) unique to the cocoa genome. This indicated an expansion of some gene families during the evolution of cocoa. Some of the genes were annotated as flavonoid related, a contributing factor to the flavour and scent of chocolate. A re-assembly of the genome with an updated version of the Newbler assembler resulted in an assembly (ICGS Assembly 1.2) with an N50 of 5.624 Mb and the largest scaffold of 18.2 Mb. The new assembly covered 84.3% of the genome.

Genetic diversity within the Rosaceae required the use of several model species as references for comparative analysis in this family. Model species identified for this purpose include strawberry (*Fragaria vesca*), peach (*Prunus persica*) and apple (Shulaev et al. 2008). Due to the complexity of the octoploid cultivated strawberry *F. × ananassa*, ($2n = 8x = 56$), the sequencing of its diploid progenitor, the woodland strawberry *F. vesca* ($2n = 2x = 14$), was undertaken (Shulaev et al. 2011). A *de novo* assembly of Roche/454, Life Technologies/SOLiD and Illumina/Solexa platform reads at 39x-combined coverage resulted in over 3,200 scaffolds with an N50 of 1.3 Mb. A total of 95% (209.8 Mb) of the genome was represented in 272 scaffolds. Resequencing at 26x coverage with Illumina validated the assembly, with 99.8% of the scaffolds and 99.98% of bases perfectly matching with Illumina reads.

Sequencing of the apple genome (Velasco et al. 2010) followed a similar approach to that used to sequence the highly heterozygous grape genome *Vitis vinifera* cv Pinot Noir (Zharkikh et al. 2008) in which a combination of paired end reads produced by Sanger sequencing and unpaired reads produced by sequencing by synthesis was shown to be an efficient way of sequencing and assembling complex heterozygous

genomes. In total, 122,146 contigs provided a 16x coverage of the 603 Mb genome. Of these 103,076 were assembled into 1,629 meta-contigs. This assembly consisted of 26 % Sanger paired end reads and 74 % 454 sequencing by synthesis paired and unpaired reads.

Cereal crops diverged from a common ancestor some 60 million years ago and whole-genome organisation exhibits a high degree of synteny (Moore et al. 1995). Rice has the smallest genome size among major cereal crops, estimated at 430 Mbp (Goff et al. 2002a) and the genome sequences of rice provide a basis for integrating and comparing biological information from rice and related cereal crops (Goff et al. 2002b; Yu et al. 2002).

The International Rice Genome Sequencing Project (IRGSP) sequenced an inbred rice cultivar, *Oryza sativa* ssp. japonica cv. Nipponbare using a clone by clone approach with BACs and PACs. 3,401 BAC and PAC clones were sequenced and assembled, resulting in a high quality reference genome anchored to a genetic map (International Rice Genome Sequencing Project 2005). Gene content analysis estimated 37,544 genes to be present, of which 17,016 were supported by 25,636 full-length cDNAs. Using this reference, analysis of rice agronomic traits was carried out on 50 wild and domesticated rice accessions (Xu et al. 2012). 6.5 million SNPs were identified and population structure analysis determined the domestication origins of rice. *Brachypodium* is a close relative of the cool season grasses and in 2006 was sequenced by the US Department of Energy Joint Genome Institute (DOE JGI) to provide a genomic bridge between rice and other agronomically important cereals (International Brachypodium Initiative 2010).

The *Sorghum bicolor* genome consists of approximately 770 Mbp in 10 chromosomes ($2n = 20$). A WGS approach was applied within the DOE-JGI community sequencing program to sequence this genome. A validation of the assembly by comparison with 27 individually sequenced BACs indicated that the assembly was 98.46 % complete with an error rate of < 1 nucleotide per 10 kb (Paterson et al. 2009). Comparison of the genome sequence with Sorghum ESTs suggests that more than 95 % of known sorghum protein-coding genes are represented in this assembly. Sequencing the sorghum genome opened opportunities for comparative studies to be carried out in the grass family between rice, sorghum, maize and *Brachypodium* (Gu et al. 2009), including the identification of conserved noncoding sequences (CNSs) between maize, rice and sorghum (Salvi et al. 2007).

Maize was domesticated about 10,000 years ago from the grass teosinte (Doebley et al. 2006). The maize genome consists of about 2.5 Gbp of DNA maintained in 10 chromosomes, which are diverse due to changes in chromatin composition as a result of an increase in long terminal repeat retrotransposons (LTR retrotransposons) (SanMiguel et al. 1998). In 2005 the NSF, the United States Department of Agriculture (USDA), and the United States Department of Energy (DOE) provided 32 million dollars to the Washington University Genome Sequencing Centre, Cold Spring Harbor, the Arizona Genome Institute and Iowa State University, to undertake a maize genome sequencing project. B73 was selected as the maize variety to be sequenced and a BAC-by-BAC approach was chosen to complement the previous maize genome sequencing assessments. The draft release of the maize genome (Schnable et al. 2009) was sequenced using a minimum tiling path of BACs (n

= 16,848) and fosmid ($n = 63$) clones derived from integrated physical, genetic and optical maps. Shotgun sequencing of clones, to 4-6x coverage was completed and sequences manually improved. From this draft sequence, more than 32,000 genes were predicted in the genome, and 99.8 % of these were found to be on reference chromosomes. The majority of the genome space (85 %) was found to contain several hundred transposable element families, spread across the genome.

The size of the wheat (*Triticum aestivum*) genome is approximately 17,000 Mbp, much larger than related cereal genomes such as barley (*Hordeum vulgare*, 5,000 Mbp), rye (*Secale cereale*, 9,100 Mbp) and oat (*Avena sativa*, 11,000 Mbp). The size and hexaploid nature of the wheat genome create significant problems in elucidating its genome sequence. The International Wheat Genome Sequencing Consortium (IWGSC) (www.wheatgenome.org) was established in 2005 to facilitate and coordinate international efforts toward obtaining the complete sequence of the bread wheat genome. The IWGSC selected the cultivar Chinese Spring as the germplasm source for the project (Gill et al. 2004). A pilot project led by the French National Institute for Agricultural Research (INRA) was initiated in 2004 to assess the BAC fingerprinting of the largest hexaploid wheat chromosome 3B, which has been shown to carry QTLs for disease resistance and wheat quality (Börner et al. 2002; Carter et al. 2012). Using flow cytometry isolated chromosomes (Kubaláková et al. 2002; Doležel et al. 2004), a total of 68,000 BAC clones of a 3B chromosome-specific BAC library (Safar et al. 2004) was fingerprinted at the French National Sequencing Centre, Genoscope, and a minimal tiling path sequenced. This successful isolation and sequencing of chromosome 3B led extensive analysis of homoeologous gene composition and evolution, diversity, recombination and the generation of a physical map for chromosome 3B (Paux et al. 2006; Paux et al. 2008; Horvath et al. 2009; Sainenac et al. 2009; Breen et al. 2010; Hao et al. 2010; Carter et al. 2012). Similarly, chromosome specific BAC libraries have been constructed for chromosomes 1D, 4D and 6D (Janda et al. 2004). To complement these activities, individual flow sorted chromosome arms are being sequenced using Illumina shotgun sequencing (Berkman et al. 2011b; Hernandez et al. 2011; Berkman et al. 2012b; Berkman et al. 2013). Ultimately, under the International Wheat Genome Sequencing Consortium, all 34 wheat chromosome arms will be sequenced (Šafář et al. 2010). While currently these efforts have not produced a finished genome, the assemblies and syntenic builds of individual chromosome arms generated by comparison with related cereals, provides access to genomic sequence for all genes, while placing the majority of genes within an approximate order and orientation. Currently, only data for chromosome 7 is publically available at www.wheatgenome.info (Lai et al. 2012a), but this resource has already provided the basis for chromosome arm specific marker discovery (Nie et al. 2012) (Table 15.1).

15.4 Next Generation Diversity Analysis

While whole genome assemblies provide the most comprehensive resource for understanding an organism, it is currently inconceivable to attempt the *de novo* assembly of each and every plant species and variant. Following the model of human diversity

Table 15.1 Below summarises the crops sequenced to date. Though not exhaustive, it gives a glimpse into the breath of application of NGS to crop research

Species Name	Reference
<i>BAC by BAC Sequencing</i>	
<i>Arabidopsis thaliana</i> Arabidopsis thaliana	The Arabidopsis Genome Initiative (2000)
<i>Cajanus cajan</i> (Pigeon pea)	Varshney et al. 2012
<i>Lotus japonicus</i> Lotus japonicus	Sato et al. 2008
<i>Medicago truncatula</i> Medicago truncatula	http://www.medicago.org/
<i>Oryza sativa ssp. japonica</i> (Nipponbare)	International Rice Genome Sequencing Project (2005)
<i>Oryza sativa ssp. japonica</i> (Nipponbare)	Barry 2001
<i>Solanum lycopersicum</i> (Tomato)	Tomato Genome Consortium 2012
<i>Whole genome shotgun sequencing</i>	
<i>Arabidopsis lyrata</i> (Rock cress)	Hu et al. 2011
<i>Brachypodium distachyon</i> Brachypodium distachyon	The International Brachypodium Initiative (2010)
<i>Brassica rapa</i> (Chiifu) (Chinese cabbage)	Wang et al. 2011
<i>Carrica papaya</i> (Papaya)	Ming et al. 2008
<i>Cicer arietinum</i> (Chickpea)	Varshney et al. 2013
<i>Citrus sinensis</i> (Sweet Orange)	http://www.phytozome.net/orange
<i>Cucumis sativus</i> (Cucumber)	Huang et al. 2009
<i>Eucalyptus grandis</i> (Eucalyptus)	Genome Network (EUCAGEN) (http://www.phytozome.net/)
<i>Fragaria vesca</i> (Woodland strawberry)	Shulaev et al. 2011
<i>Glycine max</i> (Soybean)	Schmutz et al. 2010b
<i>Linum usitatissimum</i> (Flax)	BGI (http://www.phytozome.net/)
<i>Malus x domestica</i> Borkh (Domesticated Apple)	Velasco et al. 2010
<i>Manihot esculenta</i> (Cassava)	http://www.phytozome.net/cassava
<i>Mimulus guttatus</i> (Monkey flower)	http://www.phytozome.net/mimulus
<i>Oryza sativa ssp. indica</i> (cv. 93-11 Rice)	Yu et al. 2002
<i>Oryza sativa</i> (Nipponbare)	Goff et al. 2002a
<i>Phaseolus vulgaris</i> (Common bean)	DOE-JGI (http://www.phytozome.net/commonbean)
<i>Populus trichocarpa</i> (Black cottonwood)	Tuskan et al. 2006
<i>Prunus persica</i> (Peach)	http://www.rosaceae.org/peach/genome
<i>Ricinus communis</i> (Castor bean)	Chan et al. 2010
<i>Setaria italica</i> (Foxtail Millet)	JCI (http://www.phytozome.net/foxtailmillet)
<i>Solanum tuberosum</i> (Potato)	Xu et al. 2011b
<i>Sorghum bicolor</i> (L.) Moench	Paterson et al. 2009
<i>Theobroma cacao</i> (Cocoa)	Argout et al. 2011
<i>Vitis vinifera</i> (ENTAV 115) (Grapevine)	Velasco et al. 2007
<i>Vitis vinifera</i> (PN40024) (Grapevine)	Jaillon et al. 2007
<i>Zea mays</i> (Palomero Toluqueno) (Corn)	Vielle-Calzada et al. 2009

analysis, after an initial set of diverse individuals was sequenced to provide a reference collection, the focus moved to study genome diversity using whole genome genotyping. With the rapid growth and plummeting cost of sequence data generation, the discovery, association and application of genome diversity information from NGS data is becoming increasingly attractive (Imelfort et al. 2009b; Berkman et al. 2012a). Such diversity studies using NGS data are not without challenges (Duran

et al. 2009b). These include the very large data volumes and the high error rates associated with this type of data. However, these challenges are being addressed, and NGS data mining is becoming a common approach for diversity analysis in a range of species (Seeb et al. 2011; Hayward et al. 2012b; Jiang et al. 2012; Kazakoff et al. 2012).

15.4.1 Reference-Based Diversity Analysis

Diversity analysis using a reference sequence is useful when a well-characterised and annotated genome sequence of a closely related species is available (Duran et al. 2009d). A good reference sequence would ideally be of very high quality, preferably a model organism for a particular genus. Limitations of reference-based diversity analysis include: low level of sequence coverage of the reference genome lowering the resolution and sensitivity with which variations can be identified, assembly and sequencing errors resulting from low-complexity regions and the alignment threshold used for mapping reads to reference. An alternative approach would negate the use of a reference sequence.

15.4.2 Non Reference-Based Diversity Analysis

Several approaches have been used to assess diversity between genomes without the use of a reference. These generally involve sequence comparisons of sequence reads where diversity assessment only takes into account differences between assembled, cultivar-specific reads. This approach has been implemented using transcriptome data in AutoSNP and used to accurately call SNPs (Barker et al. 2003).

15.5 SNPs

SNPs are the ultimate form of molecular genetic markers, as a nucleotide base is the smallest unit of inheritance. A SNP represents a single nucleotide difference between two individuals at a defined location. There are three different forms of SNPs: transitions (C/T or G/A), transversions (C/G, A/T, C/A, or T/G) or small insertions/deletions (indels) (Edwards et al. 2007a; Hao et al. 2011). SNPs are direct markers as the sequence information provides the exact nature of the allelic variants. Furthermore, this sequence variation can have a major impact on how the organism develops and responds to the environment. SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers near a locus of interest (Batley and Edwards 2007).

Studies of sequence diversity have recently been performed for a range of plant species. These have indicated that SNPs appear to be abundant in plant systems (Edwards et al. 2007b; Henry and Edwards 2009). SNPs are generally biallelic and only rarely triallelic. This disadvantage, when compared with multiallelic markers is compensated by their relative abundance. The low mutation rate of SNPs makes them excellent markers for the characterisation of germplasm resources (Syvanen 2001). The challenge of SNP discovery is not the identification of polymorphic nucleotide positions, but the differentiation of polymorphisms from abundant sequence errors. This is especially true for NGS data which has a higher error rate than Sanger DNA sequencing. These errors prevent the electronic mining of this data to identify potentially biologically relevant polymorphisms. A major source of sequence error comes from the balance between the need to produce the longest sequence length and the confidence that sequences are called correctly. Because of this, sequence trimming and filtering of sequence data is often performed to reduce the abundance of erroneous sequences (Kircher et al. 2011).

The identification of true polymorphisms in a background of sequence errors can be based on four methods: sequence quality values, redundancy of the polymorphism in an alignment, specificity of an allele call with a variety and co-segregation of SNPs to define a haplotype. By using the various measures of SNP confidence assessment, true SNPs may be identified with reasonable confidence from next generation DNA sequence data.

The frequency of occurrence of a polymorphism at a particular locus provides one of the best measures of confidence in the SNP representing a true polymorphism, and is referred to as the SNP redundancy score (Barker et al. 2003). By examining SNPs that have a redundancy score equal to or greater than two (two or more of the aligned sequences represent the polymorphism), the vast majority of sequencing errors are removed. True SNPs also co-segregate to define a conserved haplotype, however determining haplotypes from short-read data is challenging as sequence reads rarely include multiple SNPs. This is less of an issue for longer sequence reads from the Roche 454 system or in the application of paired reads from the Illumina or ABI SOLiD platforms.

There are many tools available for the discovery of SNPs from NGS data, but few have been designed specifically for plant populations (Appleby et al. 2009; Batley and Edwards 2009b; Duran et al. 2009b; Duran et al. 2013). One tool, based on autoSNP software (Barker et al. 2003; Batley et al. 2003) uses redundancy and haplotype co-segregation for SNP discovery. AutoSNPdb (Duran et al. 2009a) combines the SNP discovery pipeline of autoSNP with a relational database, hosting information on the polymorphisms, cultivars and gene annotations, to enable efficient mining and interrogation of the data. AutoSNPdb was originally developed for Sanger sequence data of rice, barley and Brassica (Duran et al. 2009c), but has also been applied to discover SNPs from wheat 454 data (Lai et al. 2012b) (<http://autosnpdb.appliedbioinformatics.com.au/>).

In one of the first examples of cereal SNP discovery from next generation genome sequence data, Barbazuk and co-workers identified more than 7,000 candidate SNPs between maize lines B73 and Mo17, with over 85% validation rate (Barbazuk

et al. 2007). This success is particularly impressive considering the complexity of the maize genome and the early version of Roche 454 sequencing applied, which produced an average read length of only 101 bp.

The larger data volumes from the Illumina sequencing platform enable the confident discovery of very large numbers of genome wide SNPs (Imelfort et al. 2009b; Hayward et al. 2012a; Lorenc et al. 2012). More than 1 million SNPs have been identified between six inbred maize lines (Lai et al. 2010). SNPs are more prevalent in diverse germplasm. Around 3.6 million SNPs were identified by sequencing 517 rice landraces (Huang et al. 2010). This study allowed for the association of genome variation with complex traits in rice and is a model for future studies in other species. Allen and coworkers identified 14,078 putative SNPs across representative samples of UK wheat germplasm using Illumina GAIIX sequencing of cDNA libraries (Allen et al. 2011), with a portion of these SNPs validated using KASPar assays (Orrù et al. 2009). Data production for SNP discovery from large genomes remains costly and often requires the development of consortium approaches (Edwards et al. 2012).

15.5.1 Copy Number Variation

Phenotypic diversity in plants has been attributed to differences resulting from copy number variation (CNV) and SNPs. CNV refers to differences in the number of gene loci between species or cultivars. CNVs can occur at several scales, from single genes to whole genomes. CNVs may be a consequence of previous polyploidy events and are believed to be behind the phenotypic diversity of polyploids, as shown in maize inbred lines (Springer et al. 2009). In addition, CNVs have been shown to originate from gene amplification events (Xu et al. 2011a).

CNVs may account for a greater variation in nucleotide content than SNPs and are believed to be the greatest contributing factor to genetic diversity (Redon et al. 2006; Schnable et al. 2009). Approaches to study CNV and presence-absence variation (PAV) are varied and may involve the use of oligonucleotide microarrays, mRNA or DNA sequences (Springer et al. 2009). Probes from a reference sequence are hybridised with labeled DNA from cultivars of interest and an assessment of their differential hybridisation can be performed to predict copy number. A disadvantage of this approach is that the probes from the reference sequence inherently contain variations, thus confounding analysis. Due to limitations of microarrays, CNV by sequencing (CNV-seq) has been developed (Xie and Tammi 2009). In CNV-seq, DNA fragments of a reference and sample are sequenced and mapped to a template sequence. A sliding window algorithm is then used to count copy numbers per window position. PAV analysis is a relatively new approach to examine diversity between genomes, chromosomes or regions of interest. Unlike CNV, which focuses on quantitative differences of reads between individuals, PAV seeks to identify elements uniquely present or absent in one cultivar irrespective of the gene frequency.

15.6 Perspectives

As more plant genome sequences become available and the cost of sequencing drops further, attention is shifting to the analysis, interpretation and integration of sequence data through comparative studies. The biggest challenge for highly repetitive genomes remains the resolution of low-complexity regions. Although in the future this may be addressed by emerging sequencing technologies, methods that can effectively address this limitation will greatly advance the study and analysis of large complex polyploid plant genomes. Such methods focus on complexity reduction of repetitive regions. Examples include the sequencing and analysis of low-copy regions of individual wheat chromosome arms (Berkman et al. 2011a) and consensus calling of SNPs based on coverage (Azam et al. 2012), among others. With second generation sequencing technologies becoming cheaper and producing more reads per sequencing run and longer read lengths (Berkman et al. 2012a) coupled with emerging third generation single molecule sequencing technologies which promise even longer read lengths (Lieberman et al. 2010; Rasko et al. 2011) and the increased availability of diverse reference genomes, a broader comparison of germplasm and a greater understanding of plant genome evolution over the coming years will be possible. In addition, the presentation of organism-specific databases with detailed, integrated and intuitive summaries of varied comparative analysis will become more critical and offer plant breeders with highly curated and organised reference resources.

References

- Allen AM, Barker GLA, Berry ST et al (2011) Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant Biotech J* 9:1086–1099
- Appleby N, Edwards D, Batley J (2009) New technologies for ultra-high throughput genotyping in plants. In: Somers D, Langridge P, Gustafson J (eds) *Plant Genomics*. Humana Press (USA), pp 19–40
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Argout X, Salse J, Aury JM et al (2011) The genome of *Theobroma cacao*. *Nat Genet* 43:101–108
- Azam S, Thakur V, Ruperao P et al (2012) Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (*Cicer arietinum*; Fabaceae), a crop species without a reference genome. *Am J Bot* 99:186–192
- Barbazuk WB, Emrich SJ, Chen HD et al (2007) SNP discovery via 454 transcriptome sequencing. *PLoS Biol* 5:1910–1918
- Barker G, Batley J, O’Sullivan H et al (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 19:421–422
- Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol* 125:1164–1165
- Batley J, Barker G, O’Sullivan H et al (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol* 132:84–91
- Batley J, Edwards D (2007) SNP applications in plants. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) *Association Mapping in Plants*. Springer, New York, pp 95–102

- Batley J, Edwards D (2009a) Genome sequence data: management, storage, and visualization. *Biotechniques* 46:333–336
- Batley J, Edwards D (2009b) Mining for Single Nucleotide Polymorphism (SNP) and Simple Sequence Repeat (SSR) molecular genetic markers. In: Posada D (ed) *Bioinformatics for DNA Sequence Analysis*. Humana Press (USA), pp 303–322
- Berkman PJ, Skarshewski A, Lorenc MT et al (2011a) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol J* 9:768–775
- Berkman PJ, Skarshewski A, Lorenc MT et al (2011b) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol J* 9:768–775
- Berkman PJ, Lai K, Lorenc MT, Edwards D (2012a) Next-generation sequencing applications for wheat crop improvement. *Am J Bot* 99:365–371
- Berkman PJ, Skarshewski A, Manoli S et al (2012b) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl Genet* 124:423–432
- Berkman PJ, Visendi P, Lee HC et al (2013) Dispersion and domestication shaped the genome of bread wheat. *Plant Biotechnol J*
- Bertioli DJ, Moretzsohn MC, Madsen LH et al (2009) An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. *BMC Genomics* 10:45
- Bock R (2010) The give-and-take of DNA: horizontal gene transfer in plants. *Trends Plant Sci* 15:11–22
- Börner AB, Schumann ES, Fürste AF et al (2002) Mapping of quantitative trait loci determining agronomic important characters in hexaploid wheat (*Triticum aestivum* L.). *Theor Appl Genet* 105:921–936
- Breen J, Wicker T, Kong X et al (2010) A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block. *Bmc Plant Biol* 10:98
- Brunner AM, Busov VB, Strauss SH (2004) Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends Plant Sci* 9:49–56
- Cannon SB, Sterck L, Rombauts S et al (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *Proc Natl Acad Sci U S A* 103:14959–14964
- Carter A, Garland-Campbell K, Morris C, Kidwell K (2012) Chromosomes 3B and 4D are associated with several milling and baking quality traits in a soft white spring wheat (*Triticum aestivum* L.) population. *Theor Appl Genet* 124:1079–1096
- Chan AP, Crabtree J, Zhao Q et al (2010) Draft genome sequence of the oilseed species *Ricinus communis*. *Nat Biotechnol* 28:951–956
- Consortium TTG (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641
- Cuesta-Marcos A, Szucs P, Close TJ et al (2010) Genome-wide SNPs and re-sequencing of growth habit and inflorescence genes in barley: implications for association mapping in germplasm arrays varying in size and structure. *BMC Genomics* 11:707
- Darracq A, Varre JS, Touzet P (2010) A scenario of mitochondrial genome evolution in maize based on rearrangement events. *BMC Genomics* 11:233
- Doležel J, Kubalaková M, Bartoš J, Macas J (2004) Flow cytogenetics and plant genome mapping. *Chromosome Res* 12:77–91
- Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127:1309–1321
- Dong CH, Li C, Yan XH et al (2011) Gene expression profiling of *Sinapis alba* leaves under drought stress and rewatering growth conditions with Illumina deep sequencing. *Mol Biol Rep* 39:5851–7
- Duran C, Appleby N, Clark T et al (2009a) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res* 37:D951–953
- Duran C, Appleby N, Edwards D, Batley J (2009b) Molecular genetic markers: discovery, applications, data storage and visualisation. *Curr Bioinform* 4:16–27

- Duran C, Appleby N, Vardy M et al (2009c) Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnol J* 7:326–333
- Duran C, Edwards D, Batley J (2009d) Genetic maps and the use of synteny. In: Somers D, Langridge P, Gustafson J (eds) *Plant Genomics*. Humana Press (USA), pp 41–66
- Duran C, Eales D, Marshall D et al (2010) Future tools for association mapping in crop plants. *Genome* 53:1017–1023
- Duran C, Singhania R, Raman H et al (2013) Predicting polymorphic EST-SSRs in silico. *Mol Ecol Resour* 13:538–45
- Edwards D, Forster JW, Chagné D, Batley J (2007a) What are SNPs? In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) *Association Mapping in Plants* Springer NY, pp 41–52
- Edwards D, Forster JW, Cogan NOI et al (2007b) Single Nucleotide Polymorphism Discovery. In: Oraguzie N, Rikkerink E, Gardiner S, De Silva H (eds) *Association Mapping in Plants*. Springer New York, pp 53–76
- Edwards D, Hansen D, Stajich J (2009) DNA Sequence Databases. In: Edwards D, Hanson D, Stajich J (eds) *Applied Bioinformatics*. Springer (USA), pp 1–11
- Edwards D, Batley J (2010) Plant genome sequencing: applications for crop improvement. *Plant Biotechnol J* 7:1–8
- Edwards D, Wang X (2012) Genome Sequencing Initiatives. In: Edwards D, Parkin IAP, Batley J (eds) *Genetics, Genomics and Breeding of Oilseed Brassicas*. Science Publishers Inc., New Hampshire, (USA), pp 152–157
- Edwards D, Wilcox S, Barrero RA et al (2012) Bread matters: a national initiative to profile the genetic diversity of Australian wheat. *Plant Biotechnol J* 10:703–708
- Edwards D, Batley J, Snowdon R (2013) Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet* 126:1–11
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16:77–88
- George J, Sawbridge TI, Cogan NO et al (2008) Comparison of genome structure between white clover and *Medicago truncatula* supports homoeologous group nomenclature based on conserved synteny. *Genome* 51:905–911
- Gill BS, Appels R, Botha-Oberholster AM et al (2004) A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics* 168:1087–1096
- Goff SA, Ricke D, Lan TH et al (2002a) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Sci* 296:92–100
- Goff SA, Ricke D, Lan TH et al (2002b) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Sci* 296:92–100
- Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis. *Am J Bot* 99:312–319
- Gu YQ, Ma Y, Huo N et al (2009) A BAC-based physical map of *Brachypodium distachyon* and its comparative analysis with rice and wheat. *BMC Genomics* 10:496
- Hall BG, Ehrlich GD, Hu FZ (2010) Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. *Microbiol* 156:1060–1068
- Hao C, Perretant M, Choulet F et al (2010) Genetic diversity and linkage disequilibrium studies on a 3.1-Mb genomic region of chromosome 3B in European and Asian bread wheat (*Triticum aestivum* L.) populations. *Theor Appl Genet* 121:1209–1225
- Hao Z, Li X, Xie C et al (2011) Identification of functional genetic variations underlying drought tolerance in maize using SNP markers. *J integrat plant biol* 53:641–652
- Hayward A, Dalton-Morgan J, Mason A et al (2012a) SNP discovery and applications in *Brassica napus*. *J Plant Biotechnol* (in press)
- Hayward A, Vighnesh G, Delay C et al (2012b) Second-generation sequencing for gene discovery in the Brassicaceae. *Plant Biotechnol J* 10:750–759
- Henry R, Edwards K (2009) New tools for single nucleotide polymorphism (SNP) discovery and analysis accelerating plant biotechnology. *Plant Biotechnol J* 7:311

- Hernandez P, Martis M, Dorado G et al (2011) NGS and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J* 69:377–386
- Horvath A, Didier A, Koenig J et al (2009) Analysis of diversity and linkage disequilibrium along chromosome 3B of bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* 119:1523–1537
- Hu TT, Pattyn P, Bakker EG et al (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet* 43:476–481
- Huang S, Li R, Zhang Z et al (2009) The genome of the cucumber, *Cucumis sativus* L. *Nat Genet* 41:1275–1281
- Huang XH, Wei XH, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–976
- Hyten DL, Song Q, Zhu Y et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci U S A* 103:16666–16671
- Imelfort M, Edwards D (2009) De novo sequencing of plant genomes using second-generation technologies. *Brief Bioinform* 10:609–618
- Imelfort M, Batley J, Grimmond S, Edwards D (2009a) Genome sequencing approaches and successes. In: Somers D, Langridge P, Gustafson J (eds) *Plant Genomics*. Humana Press (USA), pp 345–358
- Imelfort M, Duran C, Batley J, Edwards D (2009b) Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol J* 7:312–317
- International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467
- Janda J, Bartos J, Safar J et al (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. *Theor Appl Genet* 109:1337–1345
- Jiang Q, Yen SH, Stiller J et al (2012) Diversity Analysis of the Tree Legume *Pongamia pinnata* using PISSRs (Pongamia Inter-Simple Sequence Repeats). *J Plant Genome Sci* (in press)
- Kazakoff SH, Imelfort M, Edwards D et al (2012) Capturing the Biofuel Wellhead and Powerhouse: The Chloroplast and Mitochondrial Genomes of the Leguminous Feedstock Tree *Pongamia pinnata*. *Plos One* 7:51687
- Kim SY, Lohmueller KE, Albrechtsen A et al (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics* 12:231
- Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics* 12:382
- Kubaláková M, Vrána J, Čiháliková J et al (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* 104:1362–1372
- Lai JS, Li RQ, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42:1027–1158
- Lai K, Berkman PJ, Lorenc MT et al (2012a) WheatGenome.info: An integrated database and portal for wheat genome information. *Plant Cell Physiol* 53:1–7
- Lai K, Duran C, Berkman PJ et al (2012b) Single nucleotide polymorphism discovery from wheat next-generation sequence data. *Plant Biotechnol J* 10:743–749
- Lai K, Lorenc MT, Edwards D (2012c) Genomic databases for crop improvement. *Agronomy* 2:62–73
- Laing C, Buchanan C, Taboada EN et al (2010) Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* 11:461
- Lam HM, Xu X, Liu X et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nat Genet* 42:1053–1041
- Lee GA, Crawford GW, Liu L et al (2011a) Archaeological soybean (*Glycine max*) in East Asia: does size matter? *PLoS One* 6:26720

- Lee H, Lai K, Lorenc MT et al (2011b) Bioinformatics tools and databases for analysis of next generation sequence data. Briefings in Functional Genomics (in press)
- Lee H, Lai K, Lorenc MT et al (2012) Bioinformatics tools and databases for analysis of next generation sequence data. *Brief Funct Genomics* 2:12–24
- Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27:2987–2993
- Li R, Li Y, Zheng H et al (2010a) Building the sequence map of the human pan-genome. *Nat Biotechnol* 28:57–63
- Li YH, Li W, Zhang C et al (2010b) Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. *The New phytologist* 188:242–253
- Lieberman KR, Cherf GM, Doody MJ et al (2010) Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *J Am Chem Soc* 132:17961–17972
- Lodhi MA, Daly MJ, Ye GN et al (1995) A molecular marker based linkage map of *Vitis*. *Genome* 38:786–794
- Lorenc MT, Hayashi S, Stiller J et al (2012) Discovery of Single Nucleotide Polymorphisms in Complex Genomes Using SGSautoSNP. *Biology* 1:370–382
- Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Marshall D, Hayward A, Eales D et al (2010) Targeted identification of genomic regions using db. *Plant Methods* 6:19
- Ming R, Hou S, Feng Y et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 452:991–996
- Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal Genome Evolution – Grasses, Line up and Form a Circle. *Curr Biol* 5:737–739
- Mun J-H, Kwon S-J, Seol Y-J et al (2010) Sequence and structure of *Brassica rapa* chromosome A3. *Genome Biol* 11:94
- Nie X, Li B, Wang L et al (2012) Development of chromosome-arm-specific microsatellite markers in *Triticum aestivum* (Poaceae) using NGS technology. *Am J Bot* 99:369–371
- Orrù L, Catillo G, Napolitano F et al (2009) Characterization of a SNPs panel for meat traceability in six cattle breeds. *Food Control* 20:856–860
- Pang X, Luo H, Sun C (2012) Assessing the potential of candidate DNA barcodes for identifying non-flowering seed plants. *Plant Biol* 14:839–844
- Paterson AH, Bowers JE, Bruggmann R et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* 457:551–556
- Paux E, Roger D, Badaeva E et al (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J* 48:463–474
- Paux E, Sourdille P, Salse J et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322:101–104
- Rasko DA, Webster DR, Sahl JW et al (2011) Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med* 365:709–717
- Redon R, Ishikawa S, Fitch KR et al (2006) Global variation in copy number in the human genome. *Nature* 444:444–454
- Safar J, Bartos J, Janda J et al (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J* 39:960–968
- Šafář J, Šimková H, Kubaláková M et al (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res* 129:211–223
- Saintenac C, Falque M, Martin OC et al (2009) Detailed Recombination Studies Along Chromosome 3B Provide New Insights on Crossover Distribution in Wheat (*Triticum aestivum* L.). *Genetics* 181:393–403
- Salvi S, Sponza G, Morgante M et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Natl Acad Sci* 104:11376–11381

- SanMiguel P, Gaut BS, Tikhonov A et al (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20:43–45
- Sato S, Nakamura Y, Kaneko T et al (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res* 15:227–239
- Schlueter JA, Scheffler BE, Jackson S, Shoemaker RC (2008) Fractionation of synteny in a genomic region containing tandemly duplicated genes across glycine max, *Medicago truncatula*, and *Arabidopsis thaliana*. *J Hered* 99:390–395
- Schmutz J, Cannon SB, Schlueter J et al (2010a) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183
- Schmutz J, Cannon SB, Schlueter J et al (2010b) Genome sequence of the palaeopolyploid soybean. *Nature* 463:178–183
- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Seeb JE, Carvalho G, Hauser L et al (2011) Single-nucleotide polymorphism (SNP) discovery and applications of SNP genotyping in nonmodel organisms. *Mol Ecol Resour* 11(1):1–8
- Shibata D (2005) Genome sequencing and functional genomics approaches in tomato. *J Gen Plant Pathol* 71:1–7
- Shulaev V, Korban SS, Sosinski B et al (2008) Multiple models for Rosaceae genomics. *Plant Physiol* 147:985–1003
- Shulaev V, Sargent DJ, Crowhurst RN et al (2011) The genome of woodland strawberry (*Fragaria vesca*). *Nat Genet* 43:109–116
- Springer NM, Ying K, Fu Y et al (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet* 5:e1000734
- Syvanen AC (2001) Accessing genetic variation: Genotyping single nucleotide polymorphisms. *Nat Rev Genet* 2:930–942
- Tetz VV (2005) The pangenome concept: a unifying view of genetic information. *Med Sci Monitor* 11:HY24–29
- Tuskan GA, DiFazio SP, Teichmann T (2004) Poplar genomics is getting popular: The impact of the poplar genome project on tree research. *Plant Biol* 6:2–4
- Tuskan GA, Difazio S, Jansson S et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596–1604
- Varshney RK, Chen W, Li Y et al (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat Biotechnol* 30:83–89
- Varshney RK, Song C, Saxena RK et al (2013) Draft genome sequence of kabuli chickpea (*Cicer arietinum*): genetic structure and breeding constraints for crop improvement. *Nat Biotechnol*
- Velasco R, Zharkikh A, Troglio M et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* 2:e1326
- Velasco R, Zharkikh A, Affourtit J et al (2010) The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nat Genet* 42:833–839
- Vielle-Calzada JP, Martinez delaVO, Hernandez-Guzman G et al (2009) The Palomero genome suggests metal effects on domestication. *Science* 326:1078
- Wang X, Wang H, Wang J et al (2011) The genome of the mesopolyploid crop species *Brassica rapa*. *Nat Genet* 43:1035–1157
- Williams-Carrier R, Stiffler N, Belcher S et al (2010) Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. *Plant J* 63:167–177
- Wu DD, Zhang YP (2011) Eukaryotic origin of a metabolic pathway in virus by horizontal gene transfer. *Genomics* 98:367–369
- Xie C, Tammi MT (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* 10:80
- Xu JH, Bennetzen JL, Messing J (2011a) Dynamic Gene Copy Number Variation in Collinear Regions of Grass Genomes. *Mol Biol Evol*

- Xu X, Pan S, Cheng S et al (2011b) Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–195
- Xu X, Pan S, Cheng S et al (2011c) Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–194
- Xu X, Liu X, Ge S et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat Biotech* 30:105–111
- Young ND, Debelle F, Oldroyd GED et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature advance online publication*
- Yu J, Hu SN, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). *Science* 296:79–92
- Yue WF, Du M, Zhu MJ (2012) High Temperature in Combination with UV Irradiation Enhances Horizontal Transfer of *stx2* Gene from *E. coli* O157:H7 to Non-Pathogenic *E. coli*. *PLoS One* 7:e31308
- Zhang Z, Belcram H, Gornicki P et al (2011) Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proc Natl Acad Sci U S A* 108:18737–18742
- Zharkikh A, Troglio M, Pruss D et al (2008) Sequencing and assembly of highly heterozygous genome of *Vitis vinifera* L. cv Pinot Noir: problems and solutions. *J Biotechnol* 136:38–43

Chapter 16

Advances in Sequencing the Barley Genome

Nils Stein and Burkhard Steuernagel

Contents

16.1	Introduction	392
16.2	Next Generation Sequencing: New Perspectives for Barley Genome Analysis	392
16.2.1	Genome Survey and Genome Composition	393
16.2.2	Survey Sequencing of Sorted Chromosomes—A Gene Index of Barley	394
16.2.3	Next Generation Sequencing of BAC Clones	394
16.2.4	Genotyping by Sequencing	396
16.3	Whole Genome Sequencing of Barley—the Challenge Shifts from Sequencing to Assembly	397
16.4	Outlook	400
	References	401

Abstract Barley genome sequencing is lagging behind the status achieved for many other crop genomes although barley is ranking worldwide as fifth most important crop species. Whole genome sequencing of barley with classical Sanger sequencing technology was long meant to be too costly due to the very large genome size of more than 5 Gigabases. By the introduction of Next Generation Sequencing technology this situation has changed and fascinating new possibilities opened up for in depth barley genome analysis and whole genome sequencing. Genome composition has been revealed at unprecedented resolution. A linear gene order map comprising two thirds of all barley genes could be developed and the approach is currently adopted for other related and important cereal genomes like wheat and rye. Important technical limitations have been solved making even whole genome sequencing in barley a feasible endeavor. Provided these new possibilities, it is becoming obvious that soon sequencing per se is no longer the limiting factor but sequence assembly remains the challenge. This review will provide a brief summary of the recent developments in barley genome sequencing achieved since the introduction of Next Generation Sequencing.

Keywords Barley · Next generation sequencing · Genome sequencing · BAC clones · EST · SNP · Haplotype · Synteny · Retrotransposons · Heterochromatin · *Hordeum vulgare* · Triticeae

N. Stein (✉) · B. Steuernagel
Leibniz Institute of Plant Genetics and Crop Plant Research (IPK),
Corrensstr. 3, 06466 Gatersleben, Germany
e-mail: stein@ipk-gatersleben.de

16.1 Introduction

The barley (*Hordeum vulgare*) genome comprises over 5 billion base pairs which equals about double the size of the maize (*Zea mays*) genome or twelve times the genome of rice (*Oryza sativa*). If one assumes an overall investment of 100 million dollars for the sequencing of the rice genome it is obvious that sequencing the barley genome was not fundable as long scenarios were based on classical Sanger sequencing (Sanger et al. 1977). Until recently the main sequence information of the barley genome has been obtained by sequencing expressed genes (Expressed Sequence Tags, ESTs, Zhang et al. 2004) or contiguous stretches of genomic DNA (Bacterial Artificial Chromosome (BAC) clone contigs) selected in frame of map-based cloning of important genes (reviewed by: Stein and Graner 2004; Krattinger et al. 2009; Eversole et al. 2009). Even at this limited access to the barley genome, important features of genome organization could be revealed and this has been reviewed before (Stein 2007). More recently, due to the availability of novel high-throughput sequencing technologies (Next Generation Sequencing, NGS, Mardis 2008; Holt and Jones 2008) costs for sequencing have tremendously decreased. This has stimulated and facilitated genome sequencing and genome analysis in many organisms including the major crop species. The present chapter reviews recent achievements in barley genome analysis that were exclusively triggered by the technical innovation provided by NGS.

16.2 Next Generation Sequencing: New Perspectives for Barley Genome Analysis

NGS originally referred to novel sequencing technology concepts that provided an alternative, cost efficient platform to previous “gold-standard” Sanger sequencing (Sanger et al. 1977), which relies on fluorescent-di-desoxy-terminator chemistry in combination with capillary electrophoresis and laser-detection devices. The first commercial systems entered the market as “short read” sequencing technologies providing sequence read lengths of between 30 and 100 bp (Service 2006; Holt and Jones 2008). Keeping in mind that these sequencing technologies were especially designed for “re-sequencing” of high-quality reference genome sequences (i.e. the human genome), it was unclear how useful such technologies would prove to be for de novo sequencing of large crop genomes like barley which used to lack any kind of reference sequence. This concern was based on the difficulties of the barley genome carrying 80 % repetitive DNA (Flavell et al. 1974) with the main constituent repetitive elements expanding each over several kilobases (i.e. the BARE1 element, Manninen and Schulman 1993). De novo sequencing and assembly of such structures would instantly lead to false assemblies and gaps. It was therefore important to critically assess the true potential of short read technologies and revisit regularly if their use would impact the strategy of barley genome sequencing.

16.2.1 *Genome Survey and Genome Composition*

Two early studies were designed to test the usefulness of 30 bp (Illumina Solexa) or 100 bp (Roche/454 GS20) short read technology, respectively, for skim sequencing the barley genome (Wicker et al. 2008; Wicker et al. 2009). Sequence depth was shallow in both cases reaching to between 1 % (100 bp GS20) and 10 % (30 bp Solexa) haploid genome equivalents. At this coverage any attempt of genome sequence assembly was rather meaningless. But the statistical properties of the datasets revealed interesting characteristics of the barley genome.

In the initial study (Wicker et al. 2008) the 30 bp short reads were utilized to generate an index of mathematically defined repeats (MDR index, Kurtz et al. 2008). For this purpose all high quality bases of the almost 270 Mio sequence reads were used to count the occurrence of all 20 mers (a sequence word of 20 consecutive nucleotides). 159 million discrete 20 mers (sequence differs at least at one nucleotide from all other 20 mers) could be determined and 88 % of the discrete 20 mers occurred only once in the dataset. Almost 99 % of the discrete 20 mers occurred only between 1 and 10 times in the barley genome. One percent of the discrete 20 mers occurred eleven times or more often and thus constituted 30 % of the barley genome. One of these frequent 20 mers was present in almost 170,000 copies. Together this analysis provided a good reflection of the repetitiveness of the barley genome (Wicker et al. 2008).

In the second study the genome was sequenced to 1 % haploid genome coverage with 100 bp reads of the early Roche/454 GS20 technology platform (Wicker et al. 2009). The resulting 570,000 sequence reads were systematically compared against databases in order to determine the fraction of sequences that could either be related to genes or known repetitive DNA elements of Triticeae genomes. Interestingly but not unexpectedly, similarity to genes was detected only in less than 1 % of the entire dataset. 50 % of the sequence data could be assigned to only 14 families of transposable elements (TE) with the BARE1 family alone representing about 13 % of the barley genome. This is a 5–10 times higher frequency compared to its original copy number estimates (Manninen and Schulman 1993; Vicent et al. 1999). Sequences that could not be immediately related to any known class of DNA (TE, genes, SSR, organellar genomes) were assembled. Based on the shallow sequence depth resulting sequence clusters could per se be classified as repetitive sequence. Altogether 70 % of this “snapshot” sequence dataset comprised repetitive DNA. Based on the assumption that the dataset was representative for the overall composition of the barley genome it was compared to the sequence composition of individual BAC clone sequences from public databases. Such “gene-containing” BAC clones exhibited significantly different sequence compositions. Caspar transposons were over- and BAGY2 retrotransposons were underrepresented on the analysed set of BACs. The genome-wide distribution of these two TE classes was monitored by fluorescent in-situ hybridization (FISH) to barley metaphase spreads. The Caspar elements were predominantly clustered at the telomeric ends of chromosomes whereas BAGY2 elements produced almost a mirror image labeling of chromosomes covering all the pericentromeric regions but avoiding subtelomeric parts. This analysis of sequence element frequencies in comparison to a genomic index thus confirmed the subtelomeric origin of the analysed BAC clones (Wicker et al. 2009).

16.2.2 Survey Sequencing of Sorted Chromosomes— A Gene Index of Barley

Barley genome size is a major disadvantage for genome sequencing. On the other hand, the size of its genome is an advantage for cytogenetic applications. This feature could be exploited for PCR-based detection of markers from micro-dissected chromosomes (Sorokin et al. 1994). The first “physical” map of all barley chromosomes was produced by PCR amplification of genetic markers from micro-dissected chromosome arms or deletion chromosomes (Künzel et al. 2000). The usefulness of the size of Triticeae chromosome was further demonstrated by the feasibility of purifying mitotic chromosomes. Chromosome suspensions obtained from synchronized root tip meristems can be utilized for flow-cytometric sorting of over 90 % pure fractions of individual chromosomes (reviewed by Doležel et al. 2007). The Roche/454 GSFLX system was utilized to test whether such purified chromosomal DNA could be used for direct shotgun sequencing (Fig. 16.1). Chromosome 1H of barley was sequenced to about 1-fold coverage and the dataset provided partial sequence access to up to 80 % of all genes located on this respective chromosome (Mayer et al. 2009). By exploiting the extent of conserved synteny between barley and the sequenced genomes of rice and sorghum (International Rice Genome Sequencing Project 2005; Paterson et al. 2009), a potential linear order model of almost 2,000 barley genes detected in the shotgun sequences could be proposed (Fig. 16.1). These results stimulated a genome-wide analysis and all remaining barley chromosomes were sequenced to 1-fold coverage by using Roche/454 GSFLX Titanium (Mayer et al. 2011). The strategy previously applied to chromosome 1H was further reinforced by including into the analysis the information of a third sequenced model grass genome, *Brachypodium distachion* (The International Brachypodium Initiative 2010). A linear gene order map of more than 21,000 genes—perhaps two thirds of all barley genes—could be constructed with sequence tag access to all of these genes (Mayer et al. 2011). The strong enabling potential of survey sequencing of flow-sorted barley chromosomes was very convincing and stimulated similar studies on wheat chromosomes 1A, 1B, 1D (Wicker et al. 2011), 4A (Hernandez et al. 2011), 5A (Vitulo et al. 2011), 7BS and 7DS (Berkman et al. 2011a; Berkman et al. 2011b). The International Wheat Genome Sequencing Consortium has furthermore adopted the concept for wheat genome survey sequencing (www.wheatgenome.org).

16.2.3 Next Generation Sequencing of BAC Clones

Shotgun sequencing of larger stretches of genomic DNA (i.e. the insert of an average sized BAC clone) of barley can be challenging even if Sanger sequencing is applied. Seldom would an entire insert of a BAC immediately assemble into one single contig. This most often is hampered due to the presence of multiple copies of highly conserved members of the same class of repetitive elements. Therefore, it was uncertain how sequencing and assembly of barley BAC clones would perform on the basis of NGS. It could be shown by re-sequencing a few BACs (previously

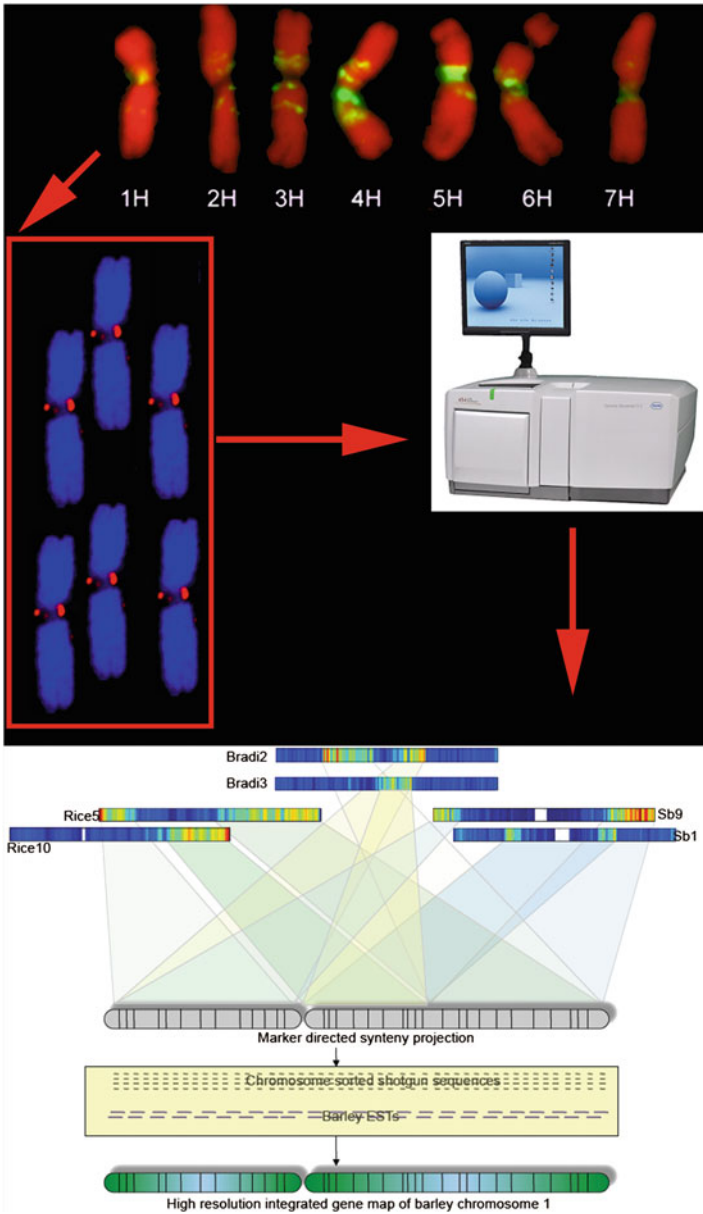


Fig. 16.1 Chromosomal genomics in barley. Barley chromosomes differ in size. This size difference can be exploited for the purification of individual chromosomes which provide, after enzymatic amplification of the minute amounts of DNA, an excellent template for Next Generation Sequencing. Sequencing to around 1-fold chromosomal coverage is providing access to about 70–80% of all genes present on the respective chromosome. This data can be integrated on the basis of a dense gene-based marker map of barley and gene order information from syntenic regions of sequenced model grass genomes into linear gene order maps of individual barley chromosomes—so-called “genome zippers” (Mayer et al. 2011; Mayer et al. 2009)

sequenced by Sanger technology) that Roche/454 GS20 sequencing could provide more even sequencing coverage due to the lack of cloning bias (Wicker et al. 2006). Assemblies were more fragmented but sequence accuracy in the assembled regions was comparable to Sanger sequencing. The even read coverage helped to detect mis-assemblies in the originally Sanger sequenced and assembled BACs. Importantly, the genes present on the sequenced BACs almost exclusively assembled into single contigs comprising the entire gene structure (Wicker et al. 2006). This feature was exploited in a recent study of sequencing 400 gene-containing BAC clones of chromosome 3H from barley cultivar Haruna Nijo. Pools of 10 or 20 BACs each were sequenced and the resulting sequences were combined in a mixed assembly. This produced 7,512 contigs larger than 500 bp. In contrast to the 444 ESTs originally used for the identification of BAC clones overall 1,239 open reading frames (ORF) could be detected in the assembled sequence and thus assigned to chromosome 3H (Sato et al. 2011).

The large sequencing capacity provided by NGS platforms per individual sequencing run requires for multiplexing of samples if the advantage of cost efficiency should not be compromised. Several procedures were established all aiming at the incorporation of unique sequence identifiers (multiplex identifier, MID; barcode tags) into the fragments obtained from different samples while ligating specific sequencing adaptors (i.e. Meyer et al. 2008). Ninety-six barley BAC clones were sequenced in parallel to average 20-fold coverage in a single Roche/454 GSFLX run. Eighty percent of the analysed clones assembled into less than 10 contigs at N50 of 50 kb for all 96 clones. Superior assembly results may be expected if BAC clones would not just be shotgun sequenced but if paired-end or mate-pair sequencing strategies would be applied which provide two sequences per DNA fragment and their physical linkage information. This strategy has been introduced in the early times of genome sequencing (Roach et al. 1995). It can be implemented today also on all NGS platforms, however, together with individual sample bar-coding it comes at the disadvantage of substantially more labor. The effect of implementing information obtained from mate-pair sequencing has been simulated for barley by combining the *de novo* shotgun 454 sequence assemblies of the above mentioned 96 BACs with additional 2×36 bp sequence reads from a non-barcoded 2.5 kb mate-pair library of the same 96 BACs which allowed to scaffold 80 % of the assembled contig length (Taudien et al. 2011). Based on the outcome of Roche/454 barcoded BAC pool sequencing two projects for sequencing chromosome 3H in barley (<http://barleygenome.org>) and 3B (<http://urgi.versailles.inra.fr/Projects/3BSeq>) in wheat have been initiated.

16.2.4 Genotyping by Sequencing

Access to high-throughput sequencing has not only stimulated research towards whole genome sequencing. It can be predicted that many of today's marker technologies will be replaced by one or the other kind of NGS application in the near/midterm future. Whole genome skim sequencing was used in rice for high-density SNP mapping in a population of 150 recombinant inbred lines (RILs). Each

individual was sequenced to average 0.02-fold genome coverage which allowed scoring of 1.2 million SNPs at a density of 3.2 SNP/kb (Huang et al. 2009). To apply this approach economically to large genome species it is appropriate to implement steps of reducing genome complexity. In restriction-associated DNA (RAD) sequencing a fraction of restriction fragments produced with specific endonucleases is size selected and sequences are generated adjacent to these restriction sites (Baird et al. 2008). A high-density haplotype map was developed by such approach for maize (Gore et al. 2009) and other species (reviewed by Rowe et al. 2011). Two studies showed the applicability to barley. 10,000 RAD fragments were generated between the parental genotypes of the Oregon Wolfe Barley (OWB) mapping population. This included 530 fragments with codominant polymorphism between both genotypes of which 436 could be genetically mapped (Chutimanitsakun et al. 2011). This approach was extended further by including presence/absence polymorphisms to the linkage analysis. For the same OWB population ~24,000 sequence tags could be mapped into the existing framework map comprising already 2,382 markers (Elshire et al. 2011).

16.3 Whole Genome Sequencing of Barley—the Challenge Shifts from Sequencing to Assembly

Whole genome sequencing in barley is feasible now because of the possibilities provided by NGS. The pure sequencing costs for the entire barley genome likely amount to less than 3 million EURO for a multiplex barcoded BACpool sequencing project with Roche/454 GS FLX+ (60,000 BACs, average insert 100 kb, 50 EURO/BAC) or to as little as 50,000 EURO for 100-fold whole genome shotgun sequencing with Illumina HiSeq2000 technology. Costs of labor and bioinformatics required for assembly and annotation have not been considered and would add up. However, having these different options in mind it is important to consider the quality of sequence that is aimed for. Recently published genome sequences produced by NGS often suffer from limited quality, hence they do allow only for limited genome wide studies and conclusions (Chain et al. 2009; Feuillet et al. 2011). The International Barley Genome Sequencing Consortium (IBSC, <http://barleygenome.org>) has proposed a multistep genome sequencing strategy which implements the advantages and quality provided by whole genome shotgun and hierarchical BAC-by-BAC sequencing utilizing a densely anchored physical map as a template (Schulte et al. 2009).

Since generating sequence data is principally no longer a limitation, the main challenge in genome sequencing shifted to sequence assembly. In any shotgun sequencing approach (WGS and hierarchical clone-by-clone shotgun sequencing) a genome needs to be reconstructed from all sequence reads by connecting overlapping reads (Miller et al. 2010). An assembly can be divided into three phases (Fig. 16.2): The first phase is the construction of contigs where a contig is a set of reads that are inter-related by overlap of their sequences (Staden 1980). All reads belong to one and only one contig. Each contig contains at least one read. In the second phase these unordered contigs are then ordered and directed. Results of the second phase are called scaffolds where a scaffold is a set of directed and ordered contigs, but gaps

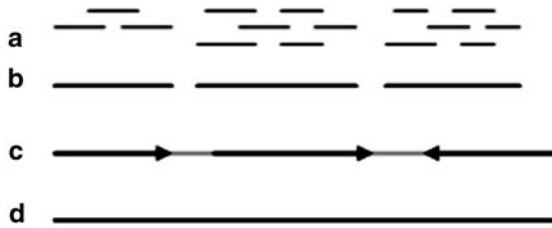


Fig. 16.2 Schematic figure of **a** reads, **b** contigs, **c** a scaffold and **d** a finished sequence. Sequence reads are provided as an output of a sequencing machine. They are used in the first phase of an assembly to produce unordered contigs. The second phase of the assembly results in scaffolds. The order and mutual direction of contigs within one scaffold is known, but gaps between the contigs are allowed. Ideally the distance of contigs is also known. In a third phase the gaps between contigs are closed

Table 16.1 Selection of current state-of-the-art NGS assembly softwares for de novo genome assembly

Assembler	DataStructure	Large Datasets ¹⁰⁾	Platforms ¹¹⁾	Comment
MIRA ¹⁾	Overlap	–	all	Very good on small but highly repetitive datasets
Newbler ²⁾	Overlap	+	454/ fasta	Distributed together with 454 Platform
CABOG ³⁾	Overlap	–	454/ fasta	Celera improved for 454
SOAPdenovo ⁴⁾	de Bruijn	+	Illumina	Used for various large genomes
CLC Assembly Cell ⁵⁾	de Bruijn	++	Illumina/ 454/ fasta	Commercial! Very efficient memory usage.
Velvet ⁶⁾	de Bruijn	–	Illumina/ fasta	Pipelines for large datasets in preparation (Cortex, Curtain)
AbySS ⁷⁾	de Bruijn	+	Illumina/ fasta	Designed for cluster computing
ALLPATH-LG ⁸⁾	de Bruijn	+	Illumina	Specific variety of paired-end-libraries required
SGA ⁹⁾	Overlap (String-graph)	++	Illumina/ 454/ fasta	Burrows-Wheeler transform for overlap detection

¹⁾ Chevreur et al. 1999, ²⁾ www.my454.com, ³⁾ Miller et al. 2010, ⁴⁾ Li et al. 2010, ⁵⁾ www.clcbio.com, ⁶⁾ Zerbino and Birney 2008, ⁷⁾ Simpson et al. 2009, ⁸⁾ Gnerre et al. 2011, ⁹⁾ Simpson and Durbin 2011, ¹⁰⁾ describes its performance on genomes larger than 2 Gb: – = not possible, + = possible on a high-end supercomputer (~1 TB RAM), ++ = possible on standard computer (~100 GB RAM); ¹¹⁾ only Illumina GAIIx and HiSeq, 454 and general fasta are concerned.

between such contigs are allowed. In a third phase, gaps between contigs within scaffolds are closed and the number of scaffolds is minimized. The result of phase three is a reference sequence ideally reflecting the sequence of the real genome.

The main challenges of whole genome sequence assembly in barley are imposed by genome size and content of repetitive DNA. Different algorithms for assembly may not necessarily cope equally well with both issues. A collection of current assembly tools for NGS data is listed in Table 16.1. Assembly softwares usually run a

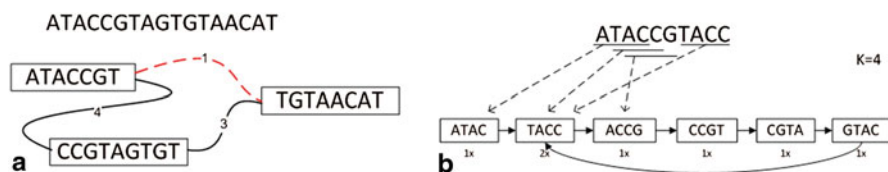


Fig. 16.3 Examples for an overlap graph **a** and a de Bruijn graph **b**. In an overlap graph every node represents one read from a sequencing machine. An edge between two nodes is drawn if the reads overlap. The edge can be weighted according to the degree of overlap. In a de Bruijn graph each occurring k -sized substring of any read is represented as a node but the substring of each node is unique. Only the number of occurrences of the substring can be tracked for each node. Thus the complexity of this graph does not increase if sequencing with a larger depth. Numbers in the graphs help to extract consensus contigs. In **a** the weight of a node represents the degree of overlap between reads. Circular structures in the graph are resolved by deleting edges with the least overlap. In **b** the numbers below each node track the number of occurrence in the input sequence. In this simple example the sequence can be constructed from the graph by following the directed graph until each node is passed as many times as its occurrence is denoted

graph-based data structure internally to compute contigs. The traditional graph structures are overlap-graphs: a node in a graph represents a sequence read and an edge between two nodes represents an overlap between the two reads (Fig. 16.3a). The overlap-graph is read-centric, containing as many nodes as reads are used. The assembled consensus contig sequence is then extracted from the graph. This may be achieved by applying weight to edges according to overlap length and include edges subsequently starting with the largest overlap and discarding every edge that would result in a circular sub-graph (Myers et al. 2000). Graph-based assembly tools like MIRA or NEWBLER proved to be efficient for highly accurate assembly of sequenced barley BAC clones (Steuernagel et al. 2009; Taudien et al. 2011; Wicker et al. 2006). At a size of 5 Gbp, barley whole genome sequencing to 50-fold coverage with 100 bp reads would produce 2.5 billion reads. Since the overlap-graph algorithms require every-read to every-other-read comparisons this results in a quadratic problem of necessary computing steps. Additionally all reasonable read-pairings have to be stored. This quickly exceeds the memory capacity of even a super-computer (~1 Terabyte RAM) and basically excludes overlap-graph based assembly algorithms for whole genome assembly in barley. This bottleneck may not occur if the construction algorithm and storage data-structure is optimized dramatically. An example for such an optimized solution is maybe provided by the optimized SGA-assembler (Simpson et al. 2009). Another alternative is provided by algorithms using a de Bruijn graph data structure (Fig. 16.3b). Here a reduction of sequence dataset complexity is achieved by extracting k -mers (a short sequence of length k , where k is a positive integer). Every node in a de Bruijn graph is a unique k -mer that occurs at least once in the input sequence dataset. An edge is always drawn between two nodes if the first node's suffix of length $k-1$ is equal to the second node's prefix of length $k-1$. Since each k -mer must only occur once in the graph, the number of nodes depends on the number of different k -mers in the genome but not on the number of input reads. This diminishes the dilemma of NGS to produce short reads and high coverage. However,

the problem to extract consensus contigs from a de Bruijn graph structure is far more complex than from an overlap graph. Tracking the number of occurrences of each k-mer in the input reads helps to resolve loops in the graph, since it is used to infer the number of occurrences of each k-mer in the consensus contigs.

Still the number of assemblers that can compute barley genome sized datasets is small. The commercial software CLC assembly cell (www.clcbio.com) is able to process such a dataset with less than 250 Gb of Memory (A standard personal computer is equipped with 4 Gb). Such an assembly of the barley genome has been made recently available providing direct access to most of the barley genes and for over 20,000 genes with transcript evidence also a genetic and physical position in the context of the barley genome physical map (The International Barley Genome Sequencing Consortium 2012). SOAP Denovo (v. 1.05) is an alternative assembler that can process barley data on a comparable computer up to a sequencing depth of 30 fold coverage (own unpublished data).

The repetitive nature of the barley genome causes the second challenge to sequence assembly since highly conserved multiple copies of repetitive DNA are a major cause of mis-assemblies. A mis-assembled sequence differs from the original DNA and mainly two classes of mis-assemblies occur. In the first case two pieces of sequence are tied together on the basis of partial sequence identity although they do not truly overlap physically. This may apply to two low copy sequences situated adjacently to highly conserved copies of the same class of TE. The second case results from collapsing two or more (almost) identical sub-sequences producing a shorter consensus sequence than in the original DNA. This situation may occur at highly conserved tandemly repeated sequences (i.e TE or genes). The main strategy of preventing repetitive DNA based mis-assemblies builds on the use of paired end sequencing. All advanced assemblers have implemented the inclusion of paired-end information for correct contig construction and most times also its incorporation for immediate scaffolding. Additionally paired-end data can be exploited to validate finished assemblies and detect mis-assemblies (Phillippy et al. 2008). These strategies could successfully be implemented in mammalian whole genome shotgun sequencing projects. The genome of giant panda (size: 2.6 Gb) was sequenced using Illumina including paired-end (and mate pair) libraries of varying insert sizes from 150 bp to 10 kb. Half of the assembled sequence could be covered with scaffolds (N50) larger than 1.3 Mb (Li et al. 2010). The assembly software ALLPATH-LG was introduced and tested on human genome shotgun data sequenced on Illumina (Gnerre et al. 2011). The software requires a specific variety of paired-end libraries reaching from overlapping read-pairs to large insert mate pairs. The N50 scaffold even reached a size of more than 11 Mb which is comparable to the quality previously obtained by Sanger-based shotgun sequencing of the human genome.

16.4 Outlook

Progress in barley genome analysis was accelerated by the availability of the different Next Generation Sequencing technology platforms. It could be demonstrated that sequencing the barley genome is feasible now due to the availability of NGS and the

decrease in sequencing costs coming along with. Access to a relatively good quality draft sequence of barley has been provided recently. How much further such draft sequences can be developed into high-quality reference sequences will be depending on, however, the introduction of a further generation of improved sequencing technology (Metzker 2010). Such platforms will allow single molecule, real time, very long read sequencing thus helping to overcome the limitations of sequence assembly due to high repetitive DNA content. Until then, the new knowledge base obtained recently for the barley genome by high-throughput NGS is providing novel opportunities and tools to the research community for addressing questions of Triticeae biology and performance much more efficiently.

References

- Baird NA, Etter PD, Atwood TS et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* 3:e3376
- Berkman P, Skarshewski A, Manoli S et al (2011a) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl Genet*. in press
- Berkman PJ, Skarshewski A, Lorenc MT et al (2011b) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol J* 9:768–775
- Chain PSG, Grafham DV, Fulton RS et al (2009) Genome project standards in a new era of sequencing. *Science* 326:236–237
- Chevreur B, Wetter T, Suhei S (1999) Genome sequence assembly using signals and additional sequence information. *Computer science and biology: proceedings of the German conference on bioinformatics (GCB)*, 99:45–56
- Chutimanitsakun Y, Nipper R, Cuesta-Marcos A et al (2011) Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. *BMC Genomics* 12:4
- Doležel J, Kubaláková M, Paux E et al (2007) Chromosome-based genomics in the cereals. *Chromosome Res* 15:51–66
- Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379
- Eversole K, Graner A, Stein N (2009) Wheat and barley genome sequencing. In: Feuillet C, Muehlbauer GJ (eds) *Genetics and genomics of the Triticeae*. Springer pp 713–742
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16:77–88
- Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet* 12:257–269
- Gnerre S, Maccallum I, Przybylski D et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci U S A* 108:1513–1518
- Gore MA, Chia J-M, Elshire RJ et al (2009) A first-generation haplotype map of maize. *Science* 326:1115–1117
- Hernandez P, Martis M, Dorado G et al (2011) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J* 69:377–386
- Holt RA, Jones SJM (2008) The new paradigm of flow cell sequencing. *Genome Res* 18:839–846
- Huang X, Feng Q, Qian Q et al (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res* 19:1068–1076

- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Krattinger S, Wicker T, Keller B (2009) Map-based cloning of genes in Triticeae (wheat and barley). In: Feuillet C, Muehlbauer GJ (eds) *Genetics and genomics of the Triticeae*. Springer pp 337–357
- Künzel G, Korzun L, Meister A (2000) Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* 154:397–412
- Kurtz S, Narechania A, Stein J, Ware D (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics* 9:517
- Li R, Fan W, Tian G et al (2010) The sequence and de novo assembly of the giant panda genome. *Nature* 463:311–317
- Manninen I, Schulman A (1993) BARE-1, a copia-like retroelement in barley (*Hordeum vulgare* L.). *Plant Mol Biol* 22:829–846
- Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133–141
- Mayer KFX, Taudien S, Martis M et al (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* 151:496–505
- Mayer KFX, Martis M, Hedley P et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 23:1249–1263
- Metzker ML (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46
- Meyer M, Stenzel U, Hofreiter M (2008) Parallel tagged sequencing on the 454 platform. *Nat Protoc* 3:267–278
- Miller JR, Koren S, Sutton G (2010) Assembly algorithms for next-generation sequencing data. *Genomics* 95:315–327
- Myers EW, Sutton GG, Delcher AL et al (2000) A whole-genome assembly of *Drosophila*. *Science* 287:2196–2204
- Paterson AH, Bowers JE, Bruggmann R et al (2009) The sorghum bicolor genome and the diversification of grasses. *Nature* 457:551–556
- Phillippy A, Schatz M, Pop M (2008) Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol* 9:R55
- Roach JC, Boysen C, Wang K, Hood L (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* 26:345–353
- Rowe HC, Renaut S, Guggisberg A (2011) RAD in the realm of next-generation sequencing technologies. *Mol Ecol* 20:3499–3502
- Sanger F, Nicklen S, Coulson A (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* 74:5463–5467
- Sato K, Motoi Y, Yamaji N, Yoshida H (2011) 454 sequencing of pooled BAC clones on chromosome 3H of barley. *BMC Genomics* 12:246
- Schulte D, Close TJ, Graner A et al (2009) The international barley sequencing consortium—at the threshold of efficient access to the barley genome. *Plant Physiol* 149:142–147
- Service RF (2006) Gene sequencing: the race for the \$1000 genome. *Science* 311:1544–1546
- Simpson JT, Durbin R (2011) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22:549–556
- Simpson J, Wong K, Jackman S et al (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19:1117–1123
- Sorokin A, Marthe F, Houben A et al (1994) Polymerase chain reaction mediated localization of RFLP clones to microisolated translocation chromosomes of barley. *Genome* 37:550–555
- Staden R (1980) A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic Acids Res* 8:3673–3694
- Stein N (2007) Triticeae genomics: advances in sequence analysis of large genome cereal crops. *Chromosome Res* 15:21–31
- Stein N, Graner A (2004) Map-based gene isolation in cereal genomes. In: Gupta P, Varshney R (eds) *Cereal genomics*. Kluwer Academic Publishers, Dordrecht, pp 331–360

- Steuernagel B, Taudien S, Gundlach H et al (2009) De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* 10:547
- Taudien S, Steuernagel B, Ariyadasa R et al (2011) Sequencing of BAC pools by different next generation sequencing platforms and strategies. *BMC Res Notes* 4:411
- The International Barley Genome Sequencing Consortium (IBSC) (2012) A physical, genetical and functional sequence assembly of the barley genome. *Nature* 491:711–716
- The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768
- Vicient CM, Suoniemi A, Anamthawat-Jonsson K et al (1999) Retrotransposon BARE-1 and its role in genome evolution in the genus *Hordeum*. *Plant Cell* 11:1769–1784
- Vitulo N, Albiero A, Forcato C et al (2011) First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PLoS One* 6:e26421
- Wicker T, Schlagenhauf E, Graner A et al (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 7:275
- Wicker T, Narechania A, Sabot F et al (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* 9:518
- Wicker T, Taudien S, Houben A et al (2009) A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J* 59:712–722
- Wicker T, Mayer KFX, Gundlach H et al (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* 23:1706–1718
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829
- Zhang H, Sreenivasulu N, Weschke W et al (2004) Large-scale analysis of the barley transcriptome based on expressed sequence tags. *Plant J* 40:276–290

Chapter 17

The Wheat Black Jack: Advances Towards Sequencing the 21 Chromosomes of Bread Wheat

Frédéric Choulet, Mario Caccamo, Jonathan Wright, Michael Alaux, Hana Šimková, Jan Šafář, Philippe Leroy, Jaroslav Doležel, Jane Rogers, Kellye Eversole and Catherine Feuillet

Contents

17.1	Introduction	406
17.2	What have we learned about the Wheat Genome so far?	407
17.2.1	The Wheat Genome: Three for One and One for All	407
17.2.2	Deciphering the Wheat Genome Composition and Organization: An ongoing Story	408
17.2.3	Tools and Technologies to Sequence, Assemble, and Annotate the Wheat Genome	413
17.2.4	Annotating the Wheat Genome	419
17.3	Strategies to Obtain a Reference Sequence of the Bread Wheat Genome	421
17.3.1	The Chromosome-Based Approach	421
17.3.2	MTP Sequencing of the 21 Wheat Chromosomes of Bread Wheat	423
17.3.3	Whole Genome Approaches can Support the Achievement of a Reference Wheat Genome Sequence	429

F. Choulet (✉) · P. Leroy · C. Feuillet
Genetics, Diversity and Ecophysiology of Cereals,
INRA Joint Research Unit 1095 Genetics, Clermont-Ferrand, France
e-mail: frederic.choulet@clermont.inra.fr

Genetics, Diversity and Ecophysiology of Cereals,
University Blaise Pascal Joint Research Unit 1095 Genetics,
Clermont-Ferrand, France

M. Caccamo · J. Wright · J. Rogers
The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, UK

M. Alaux
INRA Centre de Versailles-Grignon, Unité de Recherche en Génomique-Info,
UR 1164, Versailles, France

H. Šimková · J. Šafář · J. Doležel
Centre of the Region Haná for Biotechnological and Agricultural Research,
Institute of Experimental Botany, Olomouc, Czech Republic

K. Eversole
International Wheat Genome Sequencing Consortium, Eversole Associates,
Wyoming Road 5207, Bethesda, USA

17.4 Integration of Wheat Sequence Information in Databases	430
17.4.1 Data Integration	430
17.4.2 Wheat Databases	431
References	434

Abstract Despite its socio-economic importance and the overall recognition that a reference genome sequence has great value for crop improvement, sequencing the wheat genome has long been considered “impossible” because of the sequencing cost and bioinformatic challenges associated with the assembly of the mostly repetitive 17 Gb hexaploid genome. In the past 5 years, however, new platforms and technologies have emerged that enabled the launching of an international effort to tackle the bread wheat genome sequence using a chromosome-by-chromosome approach. In this chapter, we review the features of the wheat genome as well as the tools and technologies that can be used to sequence, assemble, and annotate a large, complex, polyploid genome. We describe the strategies and current status of the efforts towards achieving a reference sequence for the 21 chromosomes of bread wheat. Finally, we present the databases that were established to support the integration of the sequence information with other genetic and biological information.

Keywords Wheat · Polyploid · Chromosome · Flow sorting · Physical map · Genome sequence · Next generation sequencing · Assembly · Annotation · Transposable elements · Database · Data integration

17.1 Introduction

The cultivation of wheat (*Triticum* spp.) reaches far back into history as it was one of the first domesticated food crops. For more than 8,000 years, wheat has been the basic staple food of the major civilizations of Europe, West Asia, and North Africa. Today, wheat is grown on more land area (255 million hectares) than any other crop and continues to be the most important food grain source for humans (<http://www.faostat.fao.org>). With changing diets and growing world populations, rising prices for fertilizers and phytosanitary products, increasing competition between food and non-food uses, and the negative effects of high temperature and drought resulting from climate change, food supplies must double in the next few decades to meet demand (Foley et al. 2011). Already, world wheat production has not been sufficient to meet demand in 6 out of the past 10 years (<http://faostat.fao.org>). Recent studies showed that annual, worldwide yield increases have slowed between 1995 and 2005 when compared with previous years (Brisson et al. 2010). Globally, models indicate that between 1980 and 2008 wheat production declined by 5.5 % as a result of climate trends (Lobell et al. 2011). Thus, to meet the challenge of delivering safe, high-quality, and health-promoting food and feed in an economically, environmentally sensitive, and sustainable manner, a paradigm shift is needed in wheat breeding and genetics.

A combination of new tools, methods, and germplasm resources must be established for wheat to facilitate this paradigm shift. One such resource that will underpin future wheat improvement is a high quality, reference genome sequence as it will provide access to the complete gene catalogue, an unlimited amount of molecular markers to support genome-based selection of new varieties, and a framework for the efficient exploitation of natural and induced genetic diversity. In the past decade, sequencing of model plant genomes, such as those of *Arabidopsis thaliana* and rice has revolutionized our understanding of plant biology and paved the way for the development of genome sequencing projects for a number of crops (Feuillet et al. 2011). Similar advances in wheat have been hampered by the giant size of the genome, ~ 17 Gb (haploid size), due to its high content in repeats ($> 90\%$) and its hexaploid nature ($2n = 6x = 42$; 21 pairs of homologous chromosomes originating from three homoeologous sets, referred to as the A, B, and D subgenomes, each with 7 chromosome pairs).

The international collaboration on wheat genome sequencing was launched after a USDA-NSF funded workshop that confirmed the need for sequencing the wheat genome and assessed different strategies and objectives (Gill et al. 2004). These objectives were to (i) construct an accurate, sequence-ready physical map (ordered BAC contigs) of the reference hexaploid wheat (*Triticum aestivum* L.) cultivar Chinese Spring for which large genetic stocks of aneuploid lines are available, (ii) assess the feasibility of a chromosome-specific approach (i.e., constructing maps of each of the 21 chromosomes using chromosome-specific BAC libraries), and (iii) explore different strategies for gene enrichment. Following these discussions, the International Wheat Genome Sequencing Consortium (IWGSC; <http://www.wheatgenome.org>) was launched in 2005 with the aim of advancing agricultural research for wheat production and utilization by developing DNA-based tools and resources that result from the complete sequence of the common (hexaploid) wheat genome (Feuillet and Eversole 2007).

In this chapter, we present current knowledge about the composition and organization of the wheat gene and transposable element (TE) spaces, the status of technologies and methods available to date to sequence, assemble, and annotate the wheat genome, and the complementary strategies that are being developed to obtain a reference sequence of the bread wheat genome. Finally, we present the different types of databases that are currently available to access the sequence information and link it to other types of datasets to provide scientists and breeders an opportunity for an optimal exploitation of this essential information in their programs.

17.2 What have we learned about the Wheat Genome so far?

17.2.1 *The Wheat Genome: Three for One and One for All*

One of the major features of the wheat genome is a high content in TEs ($\sim 80\%$; Smith and Flavell 1975) that resulted from massive amplifications in the ancestral genome before its divergence from related species of the *Triticeae* tribe around 15 million

years ago (MYA). Individual diploid genomes from different genera (e.g. *Aegilops*, *Triticum*, *Agropyrum*) subsequently evolved independently in the past 2.5–4 MY. During the evolution of the wheat lineage, spontaneous hybridizations occurred between diploid genomes resulting in different polyploid species. Hence, the ancestral allohexaploid *Triticum aestivum* species resulted from two hybridization events that brought together the diploid A, B, and D genomes ($2n = 6x = 42$; AABBDD) (McFadden and Sears 1946; Dubcovsky and Dvorak 2007). A first hybridization occurred about 0.5 MYA between a species related to *Triticum urartu* ($2n = 2x = 14$; A^uA^u) and one or more species from the Sitopsis section, including *Aegilops speltoides* ($2n = 2x = 14$; SS) which is the closest known relative to the B genome. The resulting fertile tetraploid ($2n = 4x = 28$; AABB) was domesticated over 10,000 years ago to become known as emmer wheat (*Triticum turgidum*), eventually giving rise to the subspecies *T. turgidum* ssp. *durum*, the ancestor of the durum wheat of today, also known as pasta wheat. Some 8,000 years ago, tetraploid emmer wheat reached the region south of the Caspian Sea and hybridized with *Aegilops tauschii* ($2n = 2x = 14$), a wild diploid species with a D genome leading to a fertile hexaploid species with an AABBDD genome – the ancestral *Triticum aestivum*, or bread wheat (Zohary and Hopf 2000). Subsequent to these hybridization and polyploidization events, a number of structural and functional rearrangements resulted in genome stabilization which was accompanied by slight reductions of 2–10 % in the size of the individual homoeologous genomes compared to their diploid ancestors (Feldman and Levy 2009).

As a result of this complex history, the bread wheat genome has one of the largest and TE-rich plant genomes (40 times larger than the rice genome; IRGSP 2005) and for a long time it was considered “impossible” to sequence. The organization and composition of the wheat genome can be depicted as two main components with different evolutionary dynamics and relative importance: a small conservative part that is subjected to selection pressure and mostly corresponds to the gene space, and a much larger and more variable component which is under more dynamic evolution and comprises the TE space as well as duplicated genes and gene fragments.

17.2.2 Deciphering the Wheat Genome Composition and Organization: An ongoing Story

17.2.2.1 The Gene Space

EST and mRNA Sequencing

In 1998, a major effort on sequencing polyA-tailed transcripts was undertaken by the wheat scientific community, mostly within the framework of the International Triticeae EST Cooperative (ITEC <http://avena.pw.usda.gov/genome/>) initiative, to compensate for the lack of wheat sequences in the public databases. In total 1,073,845 ESTs were produced (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). In parallel, a major effort launched in Japan resulted in the production of 17,525

full length cDNAs (<http://trifldb.psc.riken.jp> and <http://srs.ebi.ac.uk>). cDNAs and gene-enriched genomic clones were hybridized on series of aneuploid lines lacking complete chromosomes or chromosome segments (Sears 1954; Sears and Sears 1978; Endo and Gill 1996) to locate genes along the wheat chromosomes, thereby proving first evidence of the gene space organization. A series of experiments conducted by Gill and collaborators (Gill et al. 1996a; Gill et al. 1996b; Sandhu and Gill 2002; Erayman et al. 2004) suggested that a large majority of the genes are located in a few regions (so called “gene-rich regions”) representing about 30 % of the genome. These gene-rich regions were found mainly in the distal part of the wheat chromosomes where recombination takes place. The rest of the genome, considered as “gene-poor” would be composed mostly of TEs and a few isolated genes (Sandhu and Gill 2002). Subsequently, Qi et al. (2004) mapped more than 7,000 wheat ESTs (16,099 loci) into chromosome bins and, although they confirmed previous observations on the increasing gene density towards the chromosome ends, they also found genes in proximal (centromeric) regions, suggesting a less contrasted pattern of gene distribution along the wheat chromosomes. Such experiments also enabled *in silico* comparative analyses with the rice genome sequence to study colinearity at high resolution. For example, using 4,485 wheat ESTs mapped in deletion bins, Sorrells and collaborators (Sorrells et al. 2003; La Rota and Sorrells 2004) performed a wheat-rice genome comparison and showed that wheat chromosome group 3 is the most conserved while chromosome group 5 is the least conserved compared to rice.

ESTs, however, cannot provide complete information about the wheat gene space composition. First, they do not represent the full gene content since they are limited to loci that are expressed at a level sufficient for cloning. Second, they provide only partial information on gene structure because of their limited read length (about ~550 bp on average). Finally, they provide very redundant information as many ESTs originate from highly expressed genes. A minimal gene set of 40,935 wheat unigenes has been defined by NCBI recently (<http://www.ncbi.nlm.nih.gov/unigene>) through EST clustering. However, such a clustering prevents the identification of duplicated copies of gene families (due to polyploidy or single gene duplication) and represents an underestimation of the wheat gene set. Recently, this unigene set was used to build a NimbleGen array (Rustenholz et al. 2011) and study the expression of 3,000 genes assigned to the chromosome 3B physical map. In conclusion, despite their limitations, wheat ESTs have been a useful resource for genetic and physical mapping, and for structural annotation of genes on genomic sequences.

End Sequencing of Short and Long DNA Clone Inserts

The first attempts to explore the wheat genome beyond its transcriptome were based on low-pass survey sequencing of random genomic fragments corresponding to the ends of plasmid and bacterial artificial chromosome (BAC) libraries. Analysis of 3 Mb of plasmid end sequences produced from *Ae. tauschii* (Li et al. 2004) indicated that 68 % of the sequence consisted of TEs whereas the annotation of 11 Mb of BAC end sequences (BES) from chromosome 3B of hexaploid wheat (Paux et al.

2006) revealed a repeat content of 86 %. Further, this study estimated a gene number of 6,000 for chromosome 3B which was extrapolated to suggest 36,000 genes per diploid genome in wheat. This contrasted dramatically with the predictions of Rabinowicz et al. (2005) that suggested as much as 98,000 genes for each of the three genomes of bread wheat based on the analysis of 1,597 plasmid ends from a methyl-filtration library. The results of these random analyses have provided some insights into the genome composition, but have not illuminated its organization (into gene islands for instance). Moreover, extrapolation of observations derived from such partial sequence datasets to the whole genome remains speculative. First, the insert end sequences may be biased, especially in wheat as repetitive sequences introduce bias in frequency of restriction sites that are used for cloning BAC and plasmid inserts. This was revealed during the analysis of the frequency and distribution of *Hind*III sites along 18 Mb of large sequenced contigs (see below) which showed that they were 1.5-fold overrepresented in TEs compared to the expected rate on random sequences. The bias in TE associated-*Hind*III sites that were used to build the BAC library thus resulted in an overrepresentation of TE sequences in the dataset and a bias towards repeats. Second, low-pass survey sequences cannot be assembled into sequence contigs thereby limiting the assessment of the organization of genes and TEs along the chromosomes. Finally, the limited length of sequence information precludes distinguishing pseudogenes from functional genes and discerning recently duplicated gene copies. Thus, insert-end sequencing can contribute to our knowledge of the wheat genome composition but cannot address questions related to organization such as gene and TE distribution, gene duplication, and Copy Number Variations (CNVs) that are increasingly relevant to understanding biological functions and polymorphisms associated with traits.

Individual BAC and BAC-contig Sequencing

During map-based cloning projects and comparative studies at disease resistance, storage protein, grain hardness, domestication or vernalization loci (for review see Feuillet and Salse 2009), a number of BAC clones and small BAC contigs were sequenced to obtain additional information on the composition and organization of the wheat genome. Analyses of the 3.8 Mb representing all wheat genomic sequences available in the public databases in 2005 revealed an average gene density of 1 gene per 24 kb and only ~55 % of TEs, thereby indicating a clear bias toward gene-rich regions in these first samples (Sabot et al. 2005). This was likely due to the fact that most of the regions targeted in the map-based cloning projects were telomeric and contained gene families. In contrast, sequencing randomly chosen BAC clones from the entire genome or sequencing the ends of BAC contigs representing complete chromosomes provided evidence for a more homogeneous gene distribution in the wheat genome. In a preliminary study of 4 BAC clones from a genomic BAC library of bread wheat cv. Chinese Spring, Devos et al. (2005) estimated an average gene density of 1/75 kb while Charles et al. (2008) reported a gene density of 1/100 kb homogeneously distributed after sequencing 10 BACs (1.43 Mb) randomly chosen

from a chromosome 3B-specific BAC library. Subsequently, draft sequences of 217 additional BAC clones from the Chinese Spring BAC library have been deposited in the databases (AC200765-851; AC207901-60; AC216550-85; AC232247-62; AC238983-88 and DQ767609-30) by Bennetzen and colleagues and, the automated annotation of 67 BACs (ses.library.usyd.edu.au) suggested a density of 1 gene/64 kb. While these analyses provided new insights into the organization and composition of the genome, the sampled regions represented only a minute fraction of the genome (< 0.1 %) and information remains limited to addressing TE dynamics and gene distribution as individual BAC sequences (100-150 kb) contain on average 1 or 2 genes and a few TEs most of which are sequenced only partially.

Recently, the sequencing of large Mb-sized BAC contigs from chromosome 3B (Choulet et al. 2010) and *Ae. tauschii* (Massa et al. 2011) provided novel information about the organization and content of the gene and TE spaces. In a first study, Choulet et al. (2010) sequenced and annotated 13 Mb-sized BAC contigs (18 Mb in total) selected from different regions of chromosome 3B of bread wheat cv. Chinese Spring, and performed Whole Chromosome Shotgun (WCS) sequencing. They showed that (i) genes are present along the entire chromosome and are clustered mainly (75 %) into numerous small islands of 3–4 genes separated by large blocks of TEs that are less than 1 Mb long, and (ii) genome expansion occurred homogeneously along the chromosome through specific TE bursts. In addition, the study revealed accelerated evolution through tandem or interchromosomal gene duplications in telomeric regions that led to an increase in the gene number in wheat compared to related grasses. Gene insertion activity did not disrupt dramatically the ancestral gene “backbone” but led to an increased number of non collinear genes in wheat compared to the other species. These gene rearrangements combined with the differential insertion or removal of specific TE families resulted in a contrasted sequence composition that is now observed between the proximal and distal regions of the wheat chromosomes. Based on the 148 protein coding genes identified in the 13 contigs, Choulet et al. suggested that about 8,400 genes are located on chromosome 3B and estimated a total number of 50,000 genes per diploid genome in bread wheat. In a second study, Massa et al. (2011) sequenced and annotated 9 different regions totaling 9.7 Mb of the 4 Gb D genome from the wild ancestor *Ae. tauschii* and suggested a total gene number of 36,371. Comparison with the rice, brachypodium and sorghum genomes indicated that *Ae. tauschii* had 7,813 more genes than the grass ancestral genome which was estimated to comprise 28,558 genes.

While these BAC and BAC contig analyses provided additional information about the organization and evolution of the wheat genome, the reliability of the extrapolations made from such partial data to the whole genome remains questionable. This is illustrated by the discrepancies in the estimation of gene density and total gene number observed between the different studies. The gene density estimates ranged from 1 in 75 kb in Devos et al. (2005) to about 1 in 100 kb in Charles et al. (2008) and Choulet et al. (2010), and up to 1 in 110 kb in Massa et al. (2011). All of these were higher than the 1 in 165 kb estimated by low-pass survey BAC end sequencing (Paux et al. 2006). Gene number estimates have had even larger differences ranging from

about 36,000 (Massa et al. 2011; Paux et al. 2006) to 50,000 genes (Choulet et al. 2010) and up to 98,000 genes per diploid wheat genomes (Rabinowicz et al. 2005).

17.2.2.2 The TE Space

Early studies of reannealing kinetics of single-stranded DNA fragments (Flavell et al. 1977) indicated that the wheat genome is composed of more than 83 % of repeated sequences. Staining of mitotic chromosomes with DNA fluorochromes such as DAPI, highlighted frequent heterochromatic bands that were shown to correspond to compacted chromatin containing an abundance of repeats. Estimates of the copy number of specific and highly repeated elements were obtained from sequencing and hybridization-based genome reconstruction calculations. Some retrotransposon families, such as the WIS family, were shown to be present in tens of thousands of copies (Muniz et al. 2001). Ultimately, the sequencing of individual BAC clones (Feuillet and Keller 1999; Wicker et al. 2001; Yan et al. 2003; Chantret et al. 2004), BAC end sequences (Li et al. 2004; Paux et al. 2006), and, more recently, large BAC contigs (Choulet et al. 2010; Massa et al. 2011) provided a more detailed view of the composition of the TE fraction. These studies confirmed that the TE fraction represents more than 80 % of the wheat genome with class I LTR retrotransposons comprising the majority. They also showed that less than 10 families (Fatima, Jorge, Angela, Laura, Sabrina, WIS, Wilma, and Nusif) account for > 50 % of the TE fraction and that different TE families are found in centromeric and telomeric regions (Choulet et al. 2010). The annotation of 18 Mb of large contigs of chromosome 3B (Choulet et al. 2010) revealed 3222 TEs including 800 new families indicating the potential for TE discovery in complete BAC contig sequences. These new elements were integrated into the TREP database that is dedicated to TE annotation in wheat, barley and rye (Wicker et al. 2002). In addition, the peak of transposition activity 1.4 MYA suggests that TEs have been mostly silent in wheat since that time. Interestingly, although very fragmented by nested insertions, the majority of transposable elements identified in the Mb-sized contigs were complete. Such features were missed by previous single BAC analyses because the extremities of the same element into which dozens of other TE have inserted can be as much as 200 kb apart i.e. larger than a typical BAC size (Choulet et al. 2010). These data indicate that the wheat genome expansion due to TE activity has mainly stopped and that no significant contraction has occurred yet.

Based on all of these data, a model of the organisation and evolution of the wheat genome at the chromosome level (Fig. 17.1) can be proposed. The main features of the model are (1) genes are found everywhere along the chromosomes, (2) there are no gene deserts larger than a few Mb, (3) the genes are mostly organised in small and numerous gene islands with a slightly higher density of those towards the telomeres, and (4) there are more genes and a significant number of non collinear genes than found in related grass models. While evidence for some of the mechanisms underlying the dynamics of the gene and TE spaces (e.g. TE insertions, illegitimate and unequal recombination, interchromosomal and tandem duplications) have been

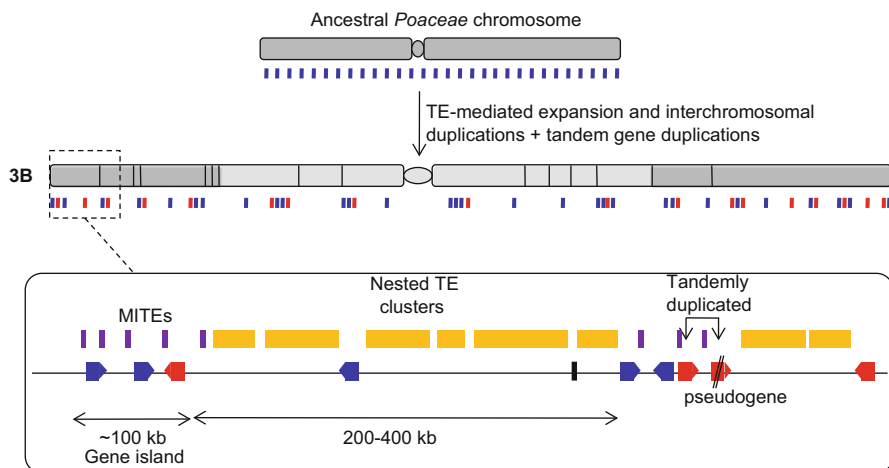


Fig. 17.1 A model for the evolution of the size, content, and organization of a wheat chromosome. Genes are represented by filled arrows: blue representing genes syntenic with other grass species and red representing locus-specific non syntenic wheat genes. TE clusters are represented by yellow rectangles while MITEs are in purple

obtained by comparative analyses of homoeologous loci (for review see Feuillet and Salse, 2009) and gene family studies (Akhunov et al. 2007), the full extent and relative impact of these mechanisms on the sequence organization remain unknown. Access to a fully annotated reference sequence will help significantly to improve our understanding of the organization and evolution of the wheat genome.

17.2.3 Tools and Technologies to Sequence, Assemble, and Annotate the Wheat Genome

17.2.3.1 Sequencing Technologies and Strategies

Prior to 2005, DNA sequencing was mostly based on the chain terminator DNA sequencing technique developed by F. Sanger (Sanger et al. 1977). Improvements in Sanger's original method that used fluorescently labeled dideoxynucleotides, automated fragment separation, and label detection facilitated the development of genomics and enabled the sequencing of entire genomes by the turn of the century (The Arabidopsis Genome Initiative 2000; Lander et al. 2001; Venter et al. 2001; Waterston et al. 2002). Sanger sequencing can produce reads of up to 1000 bases in length with low error rates, but it is relatively expensive and throughput is limited to ca. 1 million bases per day. In 2005, the emergence of the first of a second generation of sequencing technologies based on highly parallelised 'pyrosequencing' (Margulies et al. 2005) initiated a revolution in DNA sequencing and heralded the start of the competitive DNA sequencing field that exists today.

This was followed by the advent of massively parallel technologies such as the ‘sequencing-by-synthesis’ method developed by Solexa and now commercialized by Illumina (Bennett et al. 2005), and ‘Sequencing by Oligonucleotide Ligation and Detection’ (SOLiD) (Valouev et al. 2008) developed by Applied Biosystems (now Life Technologies). All of these second generation technologies rely on an amplification step to achieve the density of DNA fragments required to detect the signal of the individual bases in the DNA being sequenced. Generally, compared to Sanger sequencing, the second generation technologies produce much shorter sequencing reads, in much greater quantities, and at significantly reduced cost per sequenced base (Metzker 2009). While this reduced cost enabled the *de novo* sequencing of a large number of plant genomes (Feuillet et al. 2011), assembling large and complex repetitive genomes, such as wheat, using only short reads remains a significant challenge as the typical read lengths (100–500 bp) are not long enough to span the nested TEs that can be a few thousand bases long (Choulet et al. 2010). More recently, third-generation technologies arose that allow direct sequencing of single DNA strand and remove the dependency on an intermediate amplification step thereby reducing the execution times and potential artifacts that can arise during amplification. The Single Molecule Real Time (SRMT) technology developed by Pacific Biosciences (Eid et al. 2009) uses a polymerase attached within a nano-scale chamber to synthesize DNA complementary to a single stranded molecule attached to the polymerase within the chamber. As nucleotides are incorporated, fluorescent labels corresponding to each nucleotide are detected upon incorporation. Read lengths of up to 10 kb have been reported using this technology although reads are typically much shorter than this (1–3 kb) with the current chemistry. Another single-molecule sequencing technology is based upon detection of a change in electric current across a nanopore inserted into a lipid bilayer as individual bases are cleaved by an exonuclease from a DNA strand and pass through the nanopore (Astier et al. 2006). The identity of each base is detected by the amplitude of the change in electrical current as it passes through the nanopore. By combining high throughput and long read length, these technologies hold a great potential for sequencing large and complex genomes. Projects are underway to assess these technologies for sequencing the wheat genome (<http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=1032105>; Feuillet/Rogers/Wincker unpublished data).

The first plant genome sequencing projects of *Arabidopsis thaliana* (AGI 2000) and rice (International Rice Genome Sequencing Project 2005) were based on sequencing genome fragments cloned into vectors such as cosmids or bacterial artificial chromosomes (BACs). This is referred to as a “BAC-by-BAC” approach. In this approach, the clones are usually selected from a physical map that is assembled from the overlaps identified between BAC restriction fragment fingerprints to yield contigs representing the order of the cloned fragments in the original genome. By selecting clones from the physical map that contain fragments making up the genome sequence with minimal (~30 %) overlap between them, a minimal tiling path (MTP) of clones is defined to provide a template for sequencing. Then, each of these clones is sequenced using shotgun sub-cloning of random fragments, each clone sequence is reassembled from the shotgun reads, and the complete genome sequence is

reconstructed from the known positions and orientation of the sequenced clones. In the alternative whole genome shotgun (WGS) approach, the genome sequence is assembled from sequence reads generated from fragments of different sizes (generally ranging from 1 kb to 5 kb) obtained from the complete genome. The most effective assemblies have been built from sequence reads derived from fragments cloned into appropriately sized vectors and sequenced from both ends, yielding two reads from each clone, referred to as “mate-pairs”. As explained in more detail below, WGS sequencing depends on the development of assembly algorithms based on overlap graphs. In this approach, overlapping reads are first used to assemble regions of contiguous sequence called contigs, then mate-pair information is used to link contigs to generate longer scaffolds, thus providing a larger-scale framework for assembly of the smaller contigs of DNA (Fig. 17.2). With the emergence of the NGS technologies, the WGS approach has been established as the strategy of choice as reflected by a number of recently published genome sequences (Feuillet et al. 2011). Although the WGS approach is faster and generally cheaper, the assembly stage is much more complex increasing the risk of generating chimeric regions and misassemblies. Therefore, for sequencing large and complex plant genomes, such as wheat, the BAC-by-BAC approach is still preferred as it offers the best opportunity to generate a high quality reference sequence for the long term.

17.2.3.2 Assembling Genome Sequences

The Assembly Problem

Assembly is the reconstruction of the original genome sequence from the sequencing reads. The success of this task is limited by factors that range from the inherent complexity of genomes caused by repeat content and structure to technical issues related to specific sequencing technologies biases (e.g. error profiles and non-uniform coverage). In the general assembly pipeline, sequencing reads are assembled first into contigs, then long-range positional information, such as that obtained from mate-pairs, is used to join contigs together into scaffolds (Fig. 17.2). All assembly methods are highly dependent on the depth of coverage to which a genome is sampled, in other words, the degree of redundancy that is generated with respect to the target. For example, when enough sequence is generated to cover the target sample 10 times, the coverage is described as 10x. This sampling redundancy is essential for dealing with sequencing errors in the reads and ensures that reads will overlap sufficiently to allow accurate assembly. The first assembly algorithms were developed on the basis of the “overlap-layout-consensus approach”, i.e., the overlap information between sequencing reads is represented in a graph where the nodes are sequences and the edges connect overlapping reads. The paths in these graphs represent contigs from the underlying genome. Tools such as ARACHNE (Jaffe et al. 2003), ATLAS (Havlak et al. 2004) and more recently Newbler (<http://www.454.com>) utilize this strategy. One disadvantage of this approach is that the graph and therefore the required computer memory grow linearly with the number of reads. This is particularly

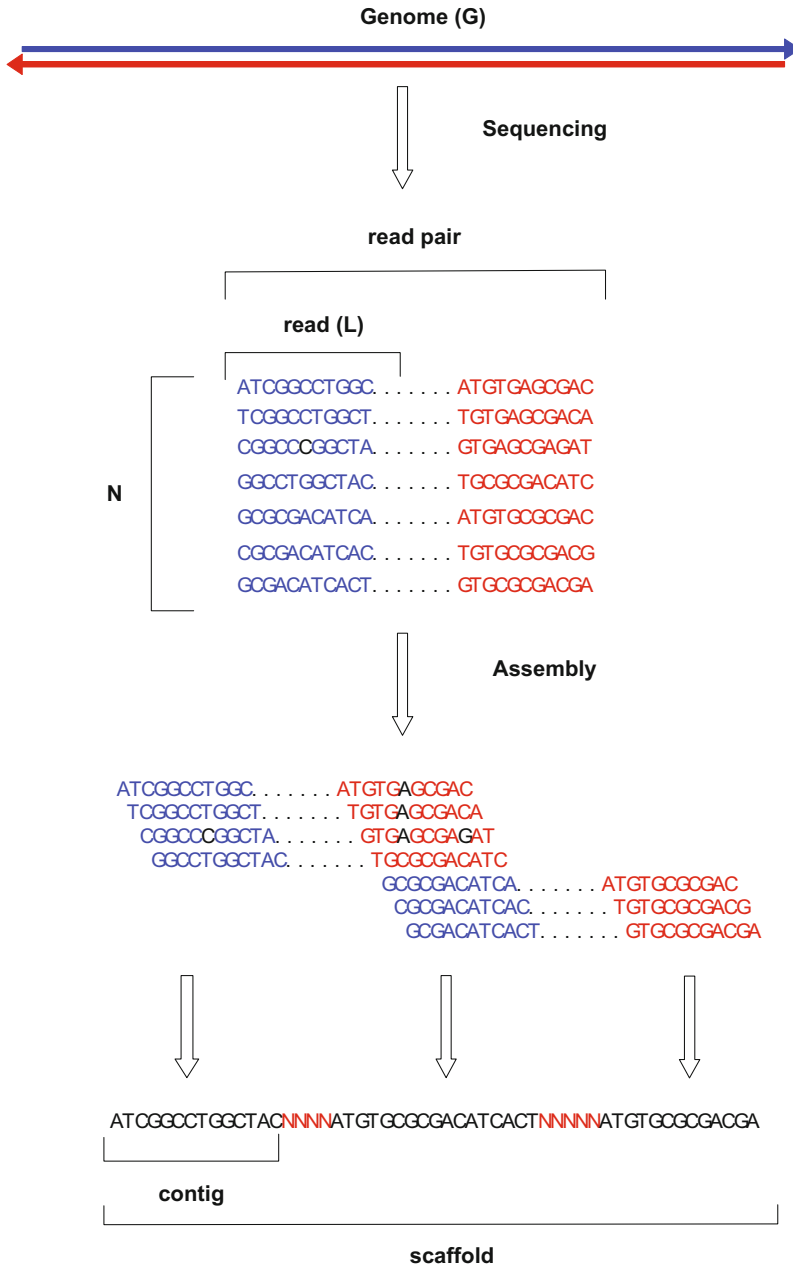


Fig. 17.2 Schematic representation of the sequence assembly process: N refers to the number of reads and L to the (average) length of the reads

relevant for NGS experiments that can generate billions of sequences per sample. The difficulty is exacerbated further when assembling large genomes as many sequencing runs are required to obtain adequate coverage. For instance, in order to generate 10x coverage of a typical 400 Mb wheat chromosome arm, 4×10^7 Illumina reads of 100 bp are required. A 10x coverage of the entire bread wheat genome represents 1.7 billion reads of 100 bp.

An alternative assembly approach pioneered by (Pevzner et al. 2001) is to represent the overlap information generated by looking only at words of fixed length k or k -mers, the so-called *de Bruijn* graphs (Fig. 17.3). The main advantage of this approach is that the graph scales with the number of observed k -mers in the dataset rather than the number of reads. In this way the graph can be built in linear time rather than the quadratic time taken with an overlap approach. In a *de Bruijn* graph, continuous stretches of nodes joined with edges represent sequence and sequencing errors appear as bubbles or tips that end abruptly (Fig. 17.3). Assembly tools generally simplify the graph by amalgamating the continuous stretches and correcting errors by bubble removal and tip clipping. Repetitive regions that share common k -mers, however, will be collapsed in single nodes generating cycles in the graph. Similarly to traditional assemblers based on sequence overlap, mate-pair reads are essential for traversing repeats in a genome and most NGS assembly algorithms incorporate this type of reads to resolve repetitive regions into scaffolds.

The Velvet assembly tool was one of the first software packages based on a *de Bruijn* graph approach to assemble short-read sequence data from bacterial genomes (Zerbino and Birney 2008). To assemble larger eukaryotic genomes, new techniques of memory optimization and parallelization were required. Algorithms such as those implemented in the open-source software packages ABySS (Simpson et al. 2009), SOAPdenovo (Li et al. 2009), ALLPATHS-LG (Gnerre et al. 2011), and Cortex (<http://cortexassembler.sourceforge.net>) allow the assembly process to be broken into multiple smaller chunks that can be distributed over multiple nodes in a high performance computing cluster. Commercial solutions are also available such as the *de novo* assembler from CLC Bio (<http://www.clcbio.com>), reputedly able to assemble a human sized genome on a single desktop computer. The SGA assembler is also designed to assemble large genomes and uses a string graph approach to determine overlaps between reads (Simpson and Durbin 2010). As well as being both memory efficient and relatively fast, SGA is the first algorithm that optimizes an overlap approach sufficiently such that it can be used to assemble large NGS datasets.

Generally, the Newbler assembler is used to assemble 454 reads as it is optimized for the error profiles and longer reads generated by the Roche platform. As such, it is being used in many of the wheat chromosome-based sequencing projects using the Roche platform such as the 3BSEQ project (<http://urgi.versailles.inra.fr/Projects/3BSeq>). In contrast, datasets generated on the Illumina platform are much larger due to the deeper coverage being generated and thus require a distributed assembly approach in addition to large amounts of computer memory. For example, as part of the IWGSC effort to obtain survey sequences of the 21 bread wheat chromosomes, the ABySS assembler is being used by The Genome Analysis Centre (<http://wheatdcc.tgac.bbsrc.ac.uk/index.php>) to produce homogeneous assemblies.

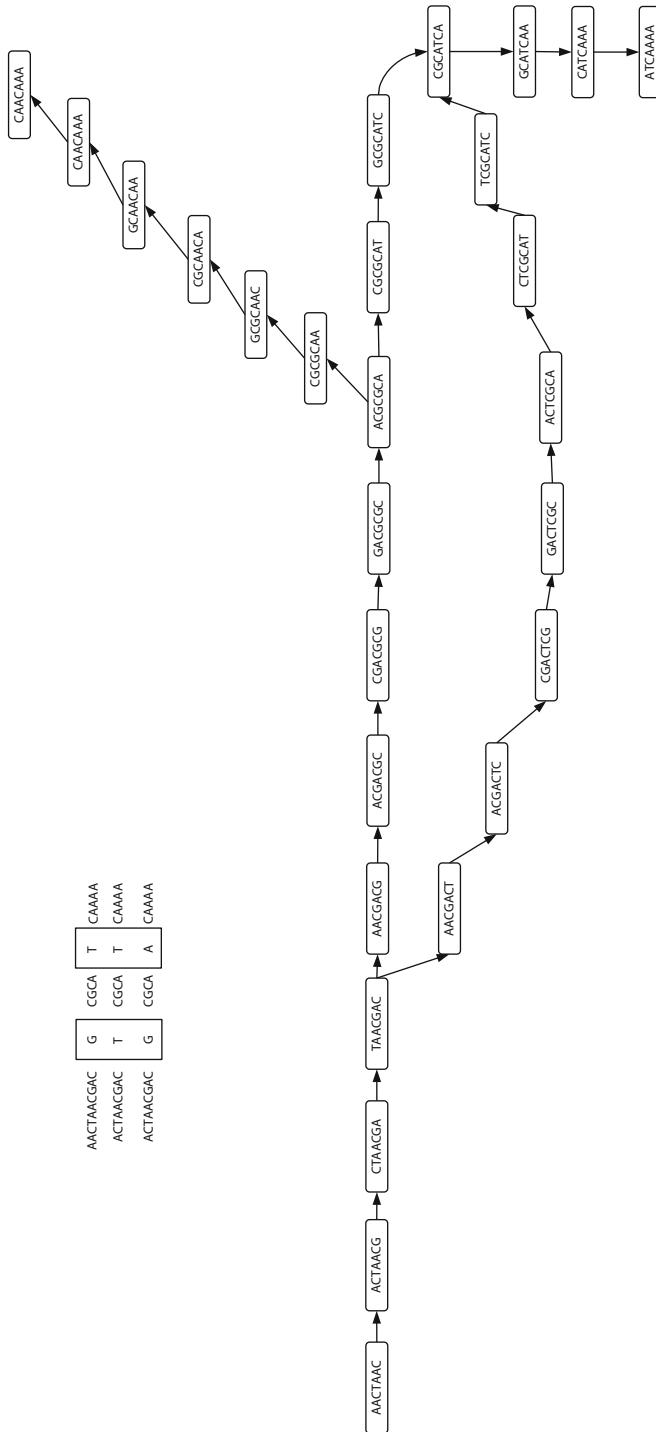


Fig. 17.3 The *de Bruijn* Graph structure. A *k-mer* length of 7 has been used to construct this graph with each *k-mer* shown as a node and joined to related nodes by edges. Errors in reads cause bubbles and tips in the graph

A Velvet-based strategy has also been adopted by other groups who have obtained comparable results on wheat chromosome group 7 (Berkman et al. 2011a, b).

17.2.3.3 Assessing the Quality of an Assembly

The N50 statistic is one of the most common metrics used to assess the quality of an assembly. After sorting the assembly contigs by size from largest to smallest, the N50 is defined as the length (L) of the contig such that half of the assembled bases are in contigs equal to or longer than L . Intuitively, the N50 represents in one figure a tradeoff between the number and median length of contigs; i.e., more complete assemblies with fewer and larger contigs will have a higher N50. The N50 statistic is further defined as ‘contig-N50’ for contigs and ‘scaffold-N50’ calculated using the scaffolds generated from the contigs. In a recent “Assemblathon” competition designed to evaluate various assembly tools on a simulated dataset (Earl et al. 2011), a NG50 statistic was introduced. Instead of using the total length of the assembled contigs as an estimation of the genome size, the NG50 used the average length of two simulated haplotypes to assess the assemblies. This allowed the authors to determine how well different assembly tools deal with haplotype-specific polymorphisms and to calculate the relative contributions of reads from each haplotype to the final assembly. Other metrics are also used to assess the quality of assembly such as sequence completeness, gene content, preservation of gene structures, and estimation of misassemblies. Generally, NGS sequencing methods are more prone to base calling errors than Sanger sequencing, thus the sequencing reads they produce are inherently “noisier” than Sanger reads. In addition to looking at the quality scores of called bases, “denoising” techniques such as k -mer frequency analysis or spectral correction alignment (Chaisson et al. 2004) can be applied to sequence data prior to assembly. The goal of these methods is to identify the k -mers (and reads) that occur at low frequency in the dataset and are likely to be erroneous as it is assumed that unique k -mers or reads should not occur with adequate sequencing depth.

17.2.4 Annotating the Wheat Genome

Achieving a robust structural and functional genome sequence annotation is essential to provide the foundation for further relevant biological studies. Genome annotation consists of identifying and attaching biological information to sequence features and it represents one of the most difficult tasks in genome sequencing projects. Genome annotation is generally a long and recursive process the difficulty of which increases with the size and complexity of the genome. It relies on a successive combination of software, algorithms, and methods, as well as the availability of accurate and updated sequence databanks. To manage the large amount of data generated by > 1 Gb genome size sequencing projects, sequence annotation needs to be automated, i.e. performed through a pipeline that combines all different programs and minimizes the subsequent long and laborious manual curation step. There are four different

categories of pipelines available. First, simple commercial softwares such as Vector NTI from invitrogen and DNASTAR from GATC can be used. However, these are not available on the web and they cannot be easily customized for specific needs. Second, suites of scripts that generate computational evidence for further manual curation have been developed in individual laboratories. For example, to annotate wheat BACs, DAWGPAWS was developed by (Estill and Bennetzen 2009) as a series of command line programs resulting in GFF output files. This type of pipeline, however, is also not available on the web and can only be used by skilled bioinformaticians. Third, “in house” automated pipelines have been developed by communities to annotate model plant genomes, e.g. rice (Ouyang and Buell 2004; IRGSP 2005) whereas major genomic resource centers such as the DOE/JGI (<http://www.phytozome.net/>), the MIPS (<http://mips.helmholtz-muenchen.de/plant/genomes.jsp>), Gramene (Liang et al. 2009), GenBank (<http://www.ncbi.nlm.nih.gov/genome/guide/build.shtml>), and the Ensembl project (Curwen et al. 2004) have developed their own pipelines to deal with multiple annotation projects. Although these pipelines are of high quality and generally are based on massive informatics resources, they are not accessible online. Finally, a number of automated annotation pipelines available on the web have been developed to assist communities in their efforts to annotate individual contigs of complete genomes. The first pipeline of this kind, RiceGAAS (Sakata et al. 2002), was developed originally for the annotation of the rice genome. Since then, a few others have been established such as DNA subway (iPlant, USA – <http://dnasubway.iplantcollaborative.org/>), FPGP (Amano et al. 2010), and MAKER (Cantarel et al. 2008). Each of these have user-friendly web interfaces; however, the online access does not enable the annotation of large genomes in a reasonable timeframe.

To support the annotation of the 17 Gb wheat genome sequence and to provide a useful resource to other communities coping with large and complex genomes, the “TriAnnot” pipeline was developed (Leroy et al. 2012). TriAnnot was conceived to: (1) enable automated, rapid, and robust structural and functional annotation of genes, transposable elements, and non-coding features; (2) be versatile, i.e. accessible through a user-friendly web interface to allow for the rapid analysis of a few hundred sequences or through a server for efficient and robust massive analysis in large scale projects (several thousand sequences); and (3) provide output files that can be retrieved and analyzed easily or that can be visualized directly on a web interface. Moreover, to ensure efficient use of the sequence information, TriAnnot enables links between the annotation and databases containing genetic and physical maps, markers, genes, QTL, phenotypes, ‘omics’ data, variomes, etc. that are available at the INRA URGI (<http://urgi.versailles.inra.fr/gnpis/>). To date, TriAnnot consists of four main panels for: (1) transposable element annotation and masking; (2) structural and functional annotation of protein-coding genes; (3) the identification of non-coding RNA genes and conserved non-coding sequences; and, (4) marker development. The performance of TriAnnot was evaluated in terms of sensitivity, specificity, and general fitness using curated reference sequence sets from rice and wheat. The results showed that TriAnnot predicted and annotated 83 and 93 % of the rice and wheat reference gene sets, respectively, with 54–67 % of those in accordance with the

reference annotations (Leroy et al. 2012). On the wheat dataset, TriAnnot demonstrated a higher fitness than three other pipelines that were not improved for wheat thereby proving its usefulness for the annotation of the wheat genome. The pipeline is accessible at <http://www.clermont.inra.fr/triannot> and is parallelized on a computing cluster that can run a 1 Gb sequence annotation in less than five days. Additional improvements such as the integration of modules to improve the annotation of TEs and non-coding RNA genes, automate the design of molecular markers, and permit online manual curation of the sequences are underway.

17.3 Strategies to Obtain a Reference Sequence of the Bread Wheat Genome

17.3.1 *The Chromosome-Based Approach*

All plant genomes sequenced so far are of diploid species and they were obtained either by sequencing BAC clones originating from the whole genome or by whole genome shotgun approaches (Feuillet et al. 2011). Because of the structural features of the wheat genome, in particular the high degree of identity between the homoeologous gene sets and the high amount of LTR transposable elements, sequencing the 21 chromosomes of wheat as a whole with the current technologies and tools would not deliver sufficient information to qualify as a reference genome sequence. Even if assembly of genic and low-copy regions is possible to some extent, such assemblies cannot be considered as finished reference genome sequences equivalent to those that have been produced for rice and *Arabidopsis*, or even the draft versions of other plant genomes (Feuillet et al. 2011). Critical information, such as recently duplicated gene families which have been shown to play a key role in speciation and non-coding intergenic sequences that also carry relevant biological information as illustrated in whole genome functional analyses of the human genome (Birney et al. 2007), would be lost in assemblies performed on the whole genome. In addition, ordering and assigning sequence scaffolds to individual chromosomes would be a daunting task.

Thus, to ensure the delivery of a reference sequence containing the information necessary for crop improvement and advanced biological studies, the IWGSC embarked on a chromosome- by-chromosome approach several years ago. The objective of this approach is to reduce the complexity of the wheat genome to manageable sizes by dealing with individual chromosomes and chromosome arms that range from 224–800 Mb and thus are comparable to already sequenced plant genomes (Dolezel et al. 2007). Previously, other strategies such as methyl filtration (Li and Gill 2004) and Cot fractionation (Lamoureux et al. 2005) have been proposed to reduce complexity and sequencing costs in wheat, however, their efficiency was too low to be applied at large scale. Unlike these strategies, the chromosome-based approach reduces sample size and complexity by sequencing smaller genome parts without sacrificing sample information content. A second important advantage of this approach is that without the

confounding effects of homoeologous sequences, the assembly of sequence reads is considerably simplified. Finally, it permits international collaboration in which chromosomes are sequenced by individual teams thereby facilitating cost sharing and rapid application into the numerous wheat breeding programs around the world.

The chromosome-based approach has been made possible due to the technological advancements obtained in the group of J. Doležel in Olomouc (Czech Republic) that enabled high-throughput chromosome isolation using flow cytometry (Vrána et al. 2000). Their approach involves preparation of liquid suspensions of intact chromosomes from root tips of hydroponically grown seedlings and classification of chromosomes stained with a DNA-specific fluorochrome according to fluorescence intensity (DNA content). Those chromosomes that can be resolved from others are then sorted. In a majority of wheat cultivars, including the IWGSC reference cultivar Chinese Spring, flow cytometric chromosome analysis (flow karyotyping) results in distribution of relative DNA content (flow karyotype) comprising three composite peaks (I – III) representing groups of chromosomes and a peak of chromosome 3B (Vrána et al. 2000). This is the only chromosome that can be purified from most of polyploid wheats (Kubaláková et al. 2002; Kubaláková et al. 2005) and thus the pioneering experiments with chromosome genomics were made with this chromosome. In some cultivars, other chromosomes than 3B can be resolved due to chromosome polymorphism (Kubaláková et al. 2002).

Cytogenetic stocks, such as telosomic and ditelosomic lines that carry chromosome arms as telocentric chromosomes (telosomes), have proven to be the most efficient method for further dissecting the polyploid wheat genome. Telosomic lines were created originally by Sears (1954) and have been used in flow cytometry to isolate chromosome arms as small as 224 Mbp (Šafář et al. 2010) which represents only 1.3 % of the genome. Of the 42 wheat chromosome arms, the long arms of chromosomes 3B (3BL) and 5B (5BL) are too large to be resolved from the peak of chromosomes 1D, 4D, and 6D. However, these two arms can be sorted as isochromosomes (chromosomes with both arms genetically identical). The complete set of telosomic lines described by Sears and Sears (1978) carry telocentric chromosomes originating from Chinese Spring with the exception of telosomes 2BS, 2DL, 4BS, 5DS, 6AL, and 7DL (designated later 7DS based on homology with 7S *Triticeae* chromosome arms) that originated from other varieties, and were backcrossed (up to 10 times) to Chinese Spring. A few of the telosomes are not truly telocentric and carry a very short second arm.

The potential utility of the chromosome-based approach depends on the sorting throughput, the purity of sorted fractions, and the quality of DNA prepared from flow-sorted chromosomes (Dolezel et al. 2004). DNA of wheat chromosomes purified according to Vrana et al. (2000) is intact and the chromosomes are suitable for preparation of high-molecular-weight DNA. The sample rate (and hence the sort rate) and the resolution of chromosome peaks on flow karyotype (and hence the purity in the sorted fractions) are inversely related. Typical rates at which wheat chromosomes are sorted range from 10–40/sec and it is realistic to sort between 200,000–400,000 chromosomes per working day using one flow sorter. For the largest chromosome 3B, this equals to 0.4–0.8 µg DNA, while for the smallest chromosome arm 1DS

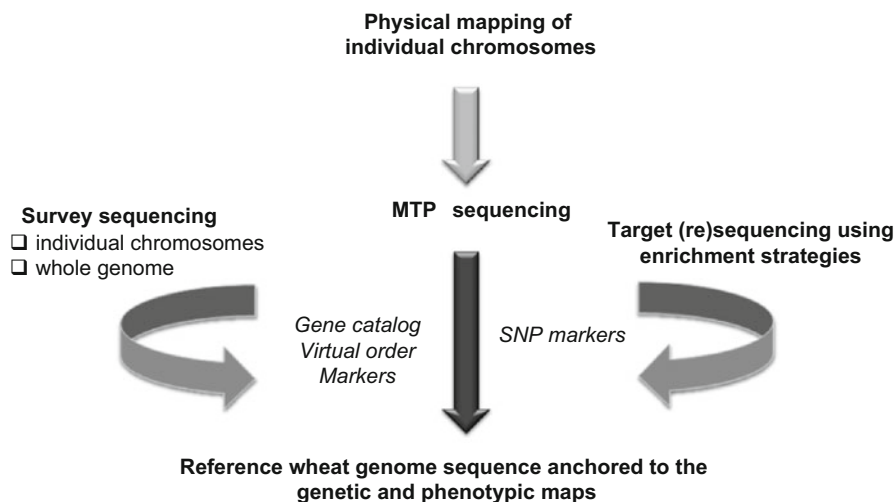


Fig. 17.4 Schematic representation of the combined strategies used by the International Wheat Genome Sequencing Consortium to obtain a reference sequence of the bread wheat genome (*T. aestivum* L. cv. Chinese Spring)

this translates to 0.09–0.18 μg DNA. The purity is best evaluated after sorting chromosomes onto a microscopic slide and fluorescent labeling of DNA repeats such as GAA microsatellites and *Afa* repeats, resulting in a chromosome-specific labeling pattern. In addition, a probe for telomeric sequence is used to confirm that intact telosomes have been sorted (Kubaláková et al. 2005). These analyses reveal that sorted wheat chromosome fractions are contaminated with fragments of chromosomes and chromatids, with no single prevailing chromosome. The purities determined during large-scale sorting experiments ranged from 80 % (telosomes 4AS, 4AL, 7AS) to 94 % (telosome 1AS). The fact that flow-sorted wheat chromosomes are intact and their DNA is of high molecular weight makes them suitable for applications ranging from BAC library construction to direct shotgun sequencing on amplified DNA (Dolezel et al. 2007).

17.3.2 MTP Sequencing of the 21 Wheat Chromosomes of Bread Wheat

17.3.2.1 Physical Mapping

In 2005, the International Wheat Genome Sequencing Consortium selected the chromosome-based approach as a foundation for its sequencing strategy for the hexaploid wheat genome (Fig. 17.4). In this strategy, physical maps are developed using chromosome-specific BAC libraries, a MTP is established for each chromosome/chromosome arm, and the BACs from the MTP are sequenced using any technology that will deliver a high quality reference sequence. To support this approach

and provide useful information (markers, gene content, etc.) while waiting for the MTP sequencing of each chromosome, whole chromosome shotgun sequences can also be produced using sorted chromosomes and combined with the physical map information (Fig. 17.4).

One of the first priorities of the IWGSC was to establish a physical map of the 21 wheat chromosomes. The physical maps, alone, facilitate the map-based isolation of genes and QTL for traits of agronomic importance and the identification of new and favorable alleles from the huge reservoir of genetic resources that are present in seed banks all over the world. Further, the development of genetically anchored physical maps can accelerate wheat improvement through the delivery of numerous markers for breeding. To reach this first priority, 37 chromosome/chromosome arm specific BAC libraries were constructed as follows: one library for chromosome 3B (Šafář et al. 2004), a composite library for chromosomes 1D, 4D, and 6D (Janda et al. 2004), and one for each arm of the remaining 34 chromosome arms. The small size of chromosome BAC libraries, typically consisting of only $3\text{--}10 \times 10^4$ clones for a chromosome coverage of $10\text{--}15 \times$ (Table 17.1), makes them easy to maintain and use (Šafář et al. 2010). As a comparison, a whole genome BAC library with similar parameters would include about 2×10^6 clones. The first chromosome-specific BAC library was constructed with only a few micrograms of chromosomal DNA from wheat chromosome 3B (Safar et al. 2004) and was used successfully to construct the first physical map of this chromosome (Paux et al. 2008), thereby validating the feasibility of constructing “sequence ready” physical maps of hexaploid wheat with a chromosome-by-chromosome approach. In total, 37 chromosome BAC libraries comprising 2,253,312 clones (Table 17.1) are available. While the use of chromosome/chromosome arm specific libraries provides important advantages, it is important to bear in mind potential limitations. The first relates to the fact that they have been constructed using a single restriction enzyme – *Hind*III (Šafář et al. 2010) which may result in uneven chromosome coverage and gaps in physical maps. The first results obtained with the 3B physical map, however, indicate 99% coverage of the chromosome (Rustenholtz et al. 2011) thereby suggesting that *Hind*III BAC libraries can provide a sufficient substrate for sequencing. One reason the coverage is sufficiently high may be related to a bias towards *Hind*III restriction sites in wheat that were observed during the sequencing of large contig sequences (Choulet et al. 2010). In fact, *Hind*III sites seem overrepresented in some of the highly repeated TE families in wheat and therefore in the whole genome. The results obtained from the first chromosome based physical maps also showed that the contamination of libraries with BAC clones originating from other chromosomes does not hamper physical map assembly as the contaminating clones are far less abundant and do not assemble into contigs. The wheat chromosome specific BAC libraries (Table 17.1) are currently being used to complete the physical map of the bread wheat genome (for a progress status as August 2013, see Fig. 17.5). Sequencing of the MTP is completed for chromosome 3B (C. Feuillet et al. unpublished) and is underway for chromosome 7B (O-A. Olsen, personal comm.).

Table 17.1 List of wheat cv. Chinese Spring chromosome-specific BAC libraries (as of January 2012)^a

Library code	Chromosome	Number of clones	kb size	Purity (%)	Coverage
TaaCsp146hA ^b	1D, 4D, 6D	87,168	85kb	91	3.4x
TaaCsp146hB ^c	1D, 4D, 6D	148,224	102kb	91	6.9x
TaaCsp146hC ^d	1D, 4D, 6D	138,240	116kb	91	7.4x
TaaCsp146eA ^{c,e}	1D, 4D, 6D	26,112	110kb	90	1.3x
TaaCsp1AShA ^c	1AS	31,104	111kb	94	11.8x
TaaCsp1ALhA ^c	1AL	49,536	103kb	83	8.0x
TaaCsp1ALhB ^d	1AL	43,008	109kb	87	7.7x
TaaCsp1BSShA ^d	1BS	55,296	113kb	81	15.7x
TaaCsp1BLhA ^d	1BL	92,160	114kb	81	15.4x
TaaCsp2AShA ^c	2AS	56,832	123kb	87	15.4x
TaaCsp2ALhA ^c	2AL	76,800	120kb	88	15.8x
TaaCsp3AShA ^b	3AS	55,296	80kb	89	10.9x
TaaCsp3AShB ^c	3AS	55,296	115kb	91	15.9x
TaaCsp3ALhA ^c	3AL	55,296	106kb	87	10.2x
TaaCsp3ALhB ^d	3AL	24,576	114kb	88	5.2x
TaaCsp3BFhA ^b	3B	67,968	103kb	89	6.2x
TaaCsp3BFhB ^c	3B	82,176	126kb	93	9.1x
TaaCsp3BFeA ^{c,e}	3B	21,120	107kb	93	1.9x
TaaCsp3DShA ^c	3DS	36,864	110kb	90	11.0x
TaaCsp3DLhA ^c	3DL	64,512	105kb	82	12.2x
TaaCsp4AShA ^d	4AS	49,152	131kb	82	16.6x
TaaCsp4ALhA ^d	4AL	92,160	126kb	81	17.3x
TaaCsp5AShA ^d	5AS	46,080	120kb	90	16.5x
TaaCsp5ALhA ^d	5AL	90,240	123kb	88	18.3x
TaaCsp5BSShA ^d	5BS	43,776	122kb	90	15.8x
TaaCsp5DShA ^d	5DS	36,864	137kb	88	17.0x
TaaCsp5DLhA ^d	5DL	72,960	128kb	87	16.0x
TaaCsp6AShA ^d	6AS	46,080	130kb	92	16.2x
TaaCsp6ALhA ^d	6AL	55,296	123 kb	86	15.7x
TaaCsp6BSShA ^d	6BS	57,600	132kb	85	15.3x
TaaCsp6BLhA ^d	6BL	76,032	130kb	92	18.0x
TaaCsp7AShA ^d	7AS	58,368	134kb	81	15.4x
TaaCsp7ALhA ^d	7AL	61,056	124kb	84	15.3x
TaaCsp7BSShA ^d	7BS	27,648	182kb	87	12.5x
TaaCsp7BLhA ^d	7BL	72,960	136kb	84	15.1x
TaaCsp7DShA ^c	7DS	49,152	114kb	84	12.2x
TaaCsp7DLhA ^c	7DL	50,304	115kb	89	14.8x
Total		2,253,312			

^aThe actual list and further details can be obtained from <http://olomouc.ueb.cas.cz/dna-libraries/cereals>

^bFirst-generation BAC libraries (one DNA size selection step)

^cSecond-generation BAC libraries (two DNA size selection steps resulting in larger inserts)

^dSecond-generation BAC libraries cloned in phage-resistant *E. coli* cells

^eBAC libraries prepared using *EcoRI* restriction endonuclease (other BAC libraries were constructed using *HindIII*)

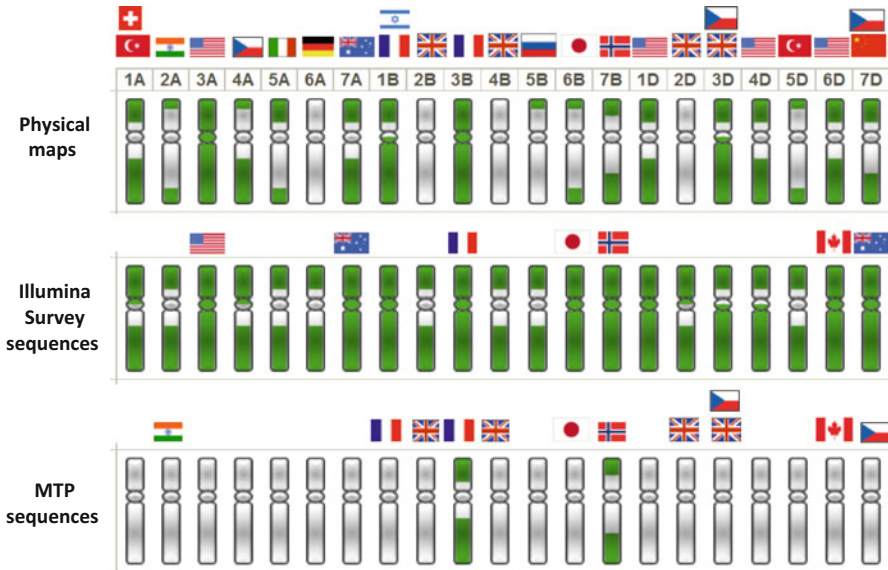


Fig. 17.5 Schematic representation of the current status (August 2013) of the IWGSC efforts to establish physical maps (a), sequence survey by Illumina (b), and reference sequences (c) of the 21 chromosomes of hexaploid wheat. The completion of each project is represented with green bars within each chromosome and chromosome arm. Four levels of completion are displayed (25, 50, 75 and 100 %). The flags represent the country of origin of the laboratories in charge of the currently funded projects. When two flags are displayed, the upper flag corresponds to the short arm and the lower flag the long arm. Except for those chromosomes with a specific flag, Illumina survey sequences were obtained by The Genome Analysis Centre (TGAC) with support from BBSRC, Biogemma, Graminor, ICARDA and INRA. For more details and regular updates, see <http://www.wheatgenome.org/Projects>

17.3.2.2 Survey Sequencing

Whole chromosome shotgun sequencing efforts began simultaneously with the production of physical maps for a number of chromosomes. It began at low-coverage to develop markers and subsequently at greater coverage to perform assemblies and gain an initial picture of the gene content. In principle, chromosomes can be sorted in adequate quantities ($> 10^6$) to obtain a sufficient amount of DNA for sequencing. However, as this would require several weeks of sorting for each of the 42 arms, DNA amplified from smaller number of sorted chromosomes ($\sim 10^4$) has been used instead. The amplification was optimized first in barley by (Simkova et al. 2008) using a commercial kit for multiple displacement amplification that achieved a highly representative amplification as validated with 1426 SNP markers. This protocol was applied to obtain 1–2x sequence coverage of barley chromosome 1H (Mayer et al. 2009) and individual arms of the remaining barley chromosomes 2H–7H (Mayer et al. 2011) using the Roche-454 technology. Integration of such low coverage shotgun sequencing information with the gene order of orthologous rice, brachypodium, and

sorghum genes (“GenomeZipper analysis”) enabled the establishment of a virtual order along the barley chromosomes for 21,766 genes (Mayer et al. 2011).

Initial survey sequencing experiments in wheat were performed for chromosome arms of the homoeologous group 1 (Wicker et al. 2011) as well as chromosome arms 7DS and 7BS (Berkman et al. 2011a, b). Wicker et al. (2011) performed Roche-454 sequencing of individual arms of the 1A, 1B, and 1D chromosomes of bread wheat and the orthologous 1H barley chromosome at low coverage (1.3–2.2 ×). At this coverage, it is not possible to perform sequence assembly, predict genes *de novo*, or distinguish genes from pseudogenes. Nevertheless, a partial gene catalog could be established for each individual chromosome arm through comparisons with sequenced genomes and used to provide for the first time an extended comparison of the gene content between 3 homoeologous wheat chromosomes. This confirmed, at the whole chromosome scale, the high proportion of non collinear genes observed previously on a subset of physical contigs for chromosome 3B (Choulet et al. 2010). Indeed, while more than 85 % of the identified genes were found in at least one chromosome, the number of non collinear genes (6158) identified for each chromosome arm largely exceeded the number of collinear genes. Moreover, the fact that 2248 nonsynthetic genes are conserved between at least two of the 1A, 1B, and 1H chromosomes suggests selection pressure and functionality for those genes. Finally, in addition to providing information about the genome composition, Roche-454 chromosome survey sequencing provided a very useful resource for designing molecular markers, such as Insertion Site Based Polymorphism (ISBP) that require read lengths of more than 200 bp to properly identify both sides of a TE junction (Paux et al. 2011).

Survey sequences of chromosome arms 7BS and 7DS (Berkman et al. 2011a, b) were produced by Illumina with coverage of 30x and 34x, respectively, using paired-end reads of 35–100 bp. The main advantages of this approach are the high coverage that can be achieved for a limited cost and the possibility of assembling the reads. The 7DS assembly resulted in 571,038 contigs accounting for only 40 % of the expected length of the chromosome arm because of collapsing reads from identical repeated elements. Comparison with markers assigned to 7DS indicated that a majority of the expected genes are present in the assembly. However, with a contig N50 value of 1.2 kb, it is likely that most genes are split into several contigs and the capacity to distinguish recently duplicated genes and pseudogenes is limited. Similar to the group 1 chromosome analysis (Wicker et al. 2011), comparative sequence studies with other grass genomes confirmed the presence of a significant proportion (31 % of the genes identified on 7DS) of non syntenic genes. In the analysis of the 7BS assembly, the position of a previously reported translocation between chromosome arms 7BS and 4AL was delimited with a resolution of one or a few genes and it was estimated that approximately 13 % of the genes have been translocated from chromosome arm 7BS to 4AL. Finally, the gene content of the 7DS and 7BS syntenic builds was used to derive an estimate of about 77,000 genes in the hexaploid wheat genome which is significantly less than the predicted 100,000–150,000 genes based on random plasmid, BAC end and BAC contig analyses (see above). These pioneering studies were followed by Roche 454 sequencing of both arms of chromosomes 5A (Vitulo et al. 2011) and 4A (Hernandez et al. 2011) further contributing to unraveling the genome structure, gene content, and gene order in hexaploid wheat.

To complete these efforts and rapidly provide chromosome-specific sequences to scientists and breeders, the IWGSC decided to systematically produce at least 50x or greater coverage for all 21 bread wheat chromosomes (<http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Sequencing/Whole-Chromosome-Survey-Sequencing>) using Illumina paired-end reads (Fig. 17.4). Although these assemblies will suffer from the same limitations as described above and will provide only partial information, they will generate a unique resource to perform *in-silico* gene and SNP marker mapping, thereby reducing the costs and labor of identifying gene locations in the hexaploid wheat genome. The sequence is generated from flow-sorted DNA for all chromosomes/chromosome arms to a depth of between 50 and 200x for each arm. For the initial stage of the project the ABySS assembler (Simpson et al. 2009) has been used to assemble the sequence reads and find the optimal *k*-mer length. Based on an assessment of the assemblies generated using various *k*-mer lengths, 71 was chosen as the optimal length to use. The N50 length of the contigs larger than 200 bp was between 1 and 4 kb (mean N50 = 2.1 kb). It varied between chromosome arms due to sequence complexity, sequencing coverage, and quality of the dataset used to generate the assembly. The order of gene-containing contigs was then inferred from synteny with the fully sequenced rice, brachypodium, and sorghum genomes using the “GenomeZipper” approach (Mayer et al. 2009). The assemblies are stored in the wheat repository at the URGI website (<http://urgi.versailles.inra.fr/Species/Wheat/Sequence-Repository>) and are publicly available.

17.3.2.3 MTP Sequencing

Typically, with an average insert size of ~120 kb, 10 BACs from a MTP are required to cover about 1 Mb of sequence. Therefore, sequencing the MTP of the 17 Gb bread wheat genome will require sequencing about 170,000 BACs. To reduce sequencing cost without losing information, BAC pooling strategies have been proposed (Rounsley et al. 2009). A reference sequence of the 1 Gb chromosome 3B is currently under production (<http://urgi.versailles.inra.fr/Projects/3BSeq>) using a pooling strategy in which 8 kb paired-end libraries are built using barcoded pools of 10 BACs from the MTP (Paux et al. 2008; Rustenholz et al. 2011) and subsequently sequenced using the Roche-454 GSFLX Titanium technology. Less than a thousand pools were needed to cover the entire 3B chromosome, and approximately 150 runs were performed to reach at minimum 30x coverage of each genomic region. A similar pooling strategy can be employed with the Illumina sequencing technology; however, the number of barcodes used simultaneously will need to be very high to optimize the coverage of each pool. Currently, HiSeq 2000 instruments can produce about 200 Gb per run. Thus, assuming that 100x coverage is sufficient to correctly assemble wheat genomic regions, 2 Gb of MTP BAC pools (i.e., 1000 pools of 2 Mb each) could theoretically be sequenced in a single run. This will require 1000 different barcodes and the capacity to ensure that DNA concentration remains equimolar in all pools to ensure equal coverage. To date,

funding has been secured for the reference MTP sequencing of 10 of the 21 wheat chromosomes (www.wheatgenome.org) and the completion of the wheat genome sequence has been recently added to the action plan on food price volatility and agriculture of the G20 within the frame of the international initiative for wheat research (http://agriculture.gouv.fr/IMG/pdf/2011-06-23_-_Action_Plan_-_VFinale.pdf).

17.3.2.4 Towards the Gold Standard Reference Sequence

Sequencing ordered MTP BAC clones will not be enough to obtain one pseudo-molecule per chromosome in a complex genome such as wheat as physical maps likely will not cover fully the entire chromosomes and small gaps will remain. Further, within the MTP contigs, not all of the sequence scaffolds will be oriented and ordered. To address some of these limitations and achieve a high quality draft sequence of chromosome 3B, a hybrid strategy that combines high coverage Illumina shotgun sequencing of DNA from sorted 3B DNA with the Roche-454 Titanium sequencing of the MTP BACs has been deployed. Sanger sequencing of BAC-ends is also being used in this strategy to facilitate the assembly of super-scaffolds for each BAC contig. Finally, markers will be designed on every sequenced super-scaffold to build high density genetic maps that will be used to order as many scaffolds as possible along the chromosome using various mapping populations (recombinants, radiation hybrid panels). While such an anchored high quality draft sequence will provide invaluable information to the scientists and breeders, additional information is needed to achieve a gold standard reference of the wheat genome. The next step change required to obtain this is a sequencing technology that can provide reads long enough to help resolve the repetitive sequences at a throughput comparable to the current short read technologies. The Single Molecule Real Time (SMRT) sequencing technology developed by Pacific Biosciences (<http://www.pacificbiosciences.com>) may offer such opportunities and pilot projects are underway to determine the extent to which this can help to achieve the reference sequence of the bread wheat genome. Finally, because of the difficulty associated with ordering the physical and sequence contigs in wheat, additional resources may be needed to reach the gold standard assembly. Optical mapping has proven very useful for validating the rice genome assembly (Zhou et al. 2007) and greatly facilitated the assembly of the maize genome (Schnable et al. 2009; Wei et al. 2009). Preliminary data indicate the suitability of DNA prepared from flow-sorted chromosomes for optical mapping (Šimková et al., unpublished) providing a potential new resource for ordering and orientating the wheat sequence scaffolds into pseudomolecules.

17.3.3 Whole Genome Approaches can Support the Achievement of a Reference Wheat Genome Sequence

While waiting for the MTP chromosome-based reference genome sequence, whole genome shotgun approaches can be useful to accelerate marker development and estimate roughly the gene content. Such WGS sequences were produced recently for

Table 17.2 Current efforts in obtaining wheat genome sequences

Species	Ploidy level	Physical map	Survey sequence	Reference sequence
<i>T. urartu</i> (A genome progenitor)	2x	Planned	Yes	No
<i>Ae. speltooides</i> (B genome related progenitor)	2x	No	Yes	No
<i>Ae. tauschii</i> (D genome progenitor)	2x	Yes	Yes	Planned
<i>T. aestivum</i>	6x			
Chromosome based ^a		Yes	Yes	Yes
Whole Genome		No	Yes	No

^asee Fig. 17.5 for a detailed representation of the efforts on each of the 21 bread wheat chromosomes

the A and D genomes of the diploid related wild species *T. urartu* and *Ae. tauschii* (Table 17.2; Jia et al. 2013; Ling et al. 2013; Brenchley et al. 2012) as well as for the hexaploid wheat cv. Chinese Spring. In the latter instance, about 200 million Roche-454 reads were produced to achieve 5x coverage of the whole genome. This sample represents the first dataset of reads homogeneously covering the whole hexaploid genome.

17.4 Integration of Wheat Sequence Information in Databases

17.4.1 Data Integration

After production, the different types of sequence data are deposited into databases. Using the model of “three-tiers of database curation” (Parkhill et al. 2010), three database categories can be distinguished: (1) non-integrated, production databases that provide access to the raw sequence data in static repositories; (2) intra-species integrated databases enabling linking genome sequences with other data (e.g. genetic and physical maps, phenotypes, markers, proteomes, etc) from the same species; and (3) inter-species integrated databases that permit viewing genome sequences in relation to data from other species. Integrated databases enable the most efficient exploitation of genomic data in biological studies. To achieve integration of data from different production databases and provide users with a unified view of these data (Lenzerini 2002), two architectures can be used: data warehouses and virtual databases (also known as federated database systems). In a data warehouse, information is offloaded from one or several production databases, aggregated, and loaded into a single database. The raw data are then cleaned, transformed, catalogued, and made available for navigation and data mining. This architecture offers a high level of data consistency with data residing together in a single repository. In addition, it ensures referential integrity, i.e. no record can be deleted if it refers to another record. A federated database system is a meta-database management system that transparently integrates multiple dispersed database systems over a computer network. Through data abstraction, federated database systems can provide a uniform interface enabling

users to manipulate data in several dispersed databases with a single operation. To this end, the system is able to decompose the query into subqueries for submission to the individual databases. Subsequently, the system merges the result sets of the subqueries into a single set. While not offering the same level of consistency as data warehouses, federated databases are easier to maintain and update.

17.4.2 *Wheat Databases*

17.4.2.1 **Non-Integrated Wheat Sequence Databases**

Wheat data are found in different non-integrated sequence databases such as the nucleotide databases of Genbank (<http://www.ncbi.nlm.nih.gov/genbank/>) and the Sequence Read Archive (SRA; <http://trace.ncbi.nlm.nih.gov/Traces/sra/>) at NCBI, the European Nucleotide Archive (ENA) at EBI (<http://www.ebi.ac.uk/ena/home>) and the DNA Databank of Japan (DDBJ; <http://www.ddbj.nig.ac.jp/>). The three organizations exchange data on a daily basis. Some linked information is available in these databases but it is very limited because it is often restricted to metadata (e.g. authors, publication reference, species taxonomy, submission date) and occasionally contains limited annotation (e.g. genes, simple repeats). Genbank provides an annotated collection of all publicly available DNA sequences while the SRA is a repository dedicated to sequence reads delivered by second generation sequencing platforms. With regard to wheat, Genbank stores all of the EST and STS sequences and SRA stores the 5x coverage survey sequences of the cv. Chinese Spring genome from CerealsDB (<http://cerealsdb.uk.net>) Table 17.3. CerealsDB also provides BLAST search facilities against an assembly of these reads and gives access to wheat SNP, EST and DArT markers. In the fall of 2011, there were 107,3845 and 1775 *Triticum aestivum* ESTs and STSs, respectively, and approximately 85 Gb of low coverage survey sequence reads in these databases. Moreover, a wheat Gene Index comprising 93,508 ESTs is available for BLAST and downloading at the Dana-Farber Cancer Institute web site (<http://compbio.dfci.harvard.edu>). In addition, several clusterings of wheat transcript sequences were performed and are available at the Unigene web site of the NCBI (<http://www.ncbi.nlm.nih.gov/unigene>) and at the portal of TIGR Plant Transcript Assemblies (<http://plantta.jcvi.org>).

17.4.2.2 **Integrated Databases for Wheat**

Wheat specific integrated databases are those dedicated to wheat “omic” and genetic data. These databases integrate the wheat sequences with other wheat data using database-constrained data links or cross-references. The level of integration varies between the databases according to the quality, nature, and quantity of the linked information.

Table 17.3 Wheat sequence databases ordered by categories and sequence types

Category	DB name	Transcript seq	Physical maps	Survey seq	Ref seq (underway)
Non-integrated	NCBI Genbank and SRA	X		X ^a	
	TIGR Plant Transcript	X			
	DFCI Gene Index	X			
	CerealsDB	X		X ^a	
Intra-species chromosome specific	WheatDB		X		
	WGGRC	X	X		
Intra-species genome-wide	Wheatgenome.info			X ^b	
	GrainGenes	X	X		
Inter-species	URGI	X	X	X ^c	X ^e
	TriFLDB	X			
Inter-species	Komugi	X			
	PlantsDB			X ^d	
	Gramene	X	X		

^alow coverage

^bhigh coverage group 7 chromosomes

^chigh coverage genome-wide

^dGenomeZipper using high coverage genome-wide

^e3B reference chromosome sequence

Chromosome-Specific Databases

With the progress in physical mapping and sequencing individual chromosomes of wheat, specific databases devoted to sets of particular chromosomes have been established Table 17.3.

The INRA URGI hosts data for the bread wheat chromosome 3B (<http://urgi.versailles.inra.fr/Species/Wheat/Data>) and for chromosome arms 3DS, 3DL, 1BS, 1BL, 1AS and 1AL as part of the database established for the European project TriticeaeGenome (<http://urgi.versailles.inra.fr/Projects/TriticeaeGenome>). Physical maps, genetic neighbor maps with links to genetic markers and QTLs as well as BAC contig sequences are available. This will be completed soon with the chromosome 3B reference sequence that is currently under analysis in the framework of the 3BSEQ project (<http://urgi.versailles.inra.fr/Projects/3BSeq>). The Wheat Genome Database at Kansas State University, WGGRC, (<http://wggrc.plantpath.ksu.edu/default.html>) provides a GBrowse access to the physical map of chromosome 3A with links to the genetic markers and BLAST facilities. It also hosts information about the physical maps of chromosomes 1D, 4D, and 6D of Chinese Spring that were developed as part of the NSF funded project on the physical mapping of the wheat D-genome (see below). WheatGenome.info (<http://www.wheatgenome.info/>) which is developed by the Australian Centre for Plant Functional Genomics and the University of Queensland provides access to the draft survey sequence reads and assemblies of bread wheat chromosomes 7A, 7B, and 7D and to a BLAST portal for these sequences.

Genome Wide Databases

GrainGenes (<http://wheat.pw.usda.gov/>) was built to provide a suite of services for the *Triticeae* and oat communities, including databases, documents, tools, data files, announcements, curation, and community assistance Table 17.3. To date, GrainGenes stores 76 wheat genetic maps, more than 100,000 genetic markers, and approximately 271,000 wheat ESTs. These sequences can be searched through a BLAST server or by using queries to get additional information on genetic mapping data. GrainGenes also hosts the Triticeae Repeat (TREP) databank that comprises 1717 sequences of wheat transposable elements. As GrainGenes is dedicated to Triticeae and Avenae species, it presents some features of inter-species integrated database, including a GBrowse display of wheat, barley, and oat EST sequences mapped on the rice genome.

A specific website (<http://avena.pw.usda.gov/wheatD/>) is dedicated to the US National Science Foundation funded physical mapping project of *Ae. tauschii*, the bread wheat D genome progenitor. It provides links to physical mapping data and enables BLAST searches against survey sequences of *Ae. tauschii*.

The INRA URGI Wheat database (<http://urgi.versailles.inra.fr/Species/Wheat>) stores 26 wheat genetic maps, 19,029 markers, 324 QTLs, 10,819 SNPs and 544,529 ESTs Table 17.3. A GBrowse is available to display physical maps in relation with other datasets (e.g., genetic markers, reference sequences, QTLs and SNPs). Physical maps of cv. Chinese Spring chromosomes that are constructed under the framework of the IWGSC are being integrated regularly into the database (<http://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Physical-mapping/>). To date, physical maps of chromosomes 1BL, 1AS, 3B, and 3DS have been integrated and a link to the 3A physical map at WGGRC is provided. In addition, URGI hosts the IWGSC sequence repository that provides access to the survey sequence assemblies of the 21 chromosomes of Chinese Spring (<http://urgi.versailles.inra.fr/Species/Wheat/Sequence-Repository>).

17.4.2.3 Inter-Species Sequence Integrated Databases

Inter-species integrated databases are designed to compare wheat sequences with related species sequences. The first way of performing inter-species sequence integration is through sequence comparisons that display the results as a repository with flat files to download or BLAST. For example, TriFLDB (<http://trifldb.psc.riken.jp>) provides access for download and BLAST against 8530 and 7341 full-length cDNAs of wheat and barley, respectively. These full-length cDNAs are further integrated into genome browsers of rice and sorghum. The second method for inter-species integration is to display the results in textual or graphical web pages that enable adding a limited level of intra-species integration when needed. The Komugi database (<http://www.shigen.nig.ac.jp>) has established a BLAST server on wheat ESTs and full length cDNAs, as well as a comparative map tool to display homologies between wheat and barley genic sequences along the rice chromosomes (<http://earth.lab.nig.ac.jp>).

PlantsDB developed by the Munich Information Centre for Protein Sequences (<http://mips.helmholtz-muenchen.de/plant/index.jsp>) has a tool called “GenomeZipper” that derives a putative chromosomal gene order for one species on the basis of its syntenic relationship with related species. Each syntenic gene presents additional information about genetic and physical maps and genomic annotation using a browser. Additional information that links the syntenic genes to the genetic and physical maps and to their annotation is available through a browser. A barley GenomeZipper is available already and a wheat GenomeZipper is underway in the framework of the IWGSC Survey Sequencing Initiative.

Gramene (<http://www.gramene.org/>) stores a number of wheat data (e.g. markers, genes) with their alignments to other cereal crops. In collaboration with EBI, efforts are underway to develop EnsemblPlants (<http://plants.ensembl.org>) which will provide access to individual genome data as well as comparative tools, such as Plant Compara. To date, EnsemblPlants does not contain any wheat sequences, but future genome sequences will likely be incorporated as they are completed.

Acknowledgments The authors want to thank Hadi Quesneville, Daphné Verdelet, Kirsley Chenen for their feedback on wheat databases. H.Š., J.S. and J.D. are supported by the Ministry of Education, Youth and Sports of the Czech Republic, the European Regional Development Fund (Operational Programme Research and Development for Innovations No. ED0007/01/01) and by the Czech Science Foundation (award no. P501/10/1740). F. C., P. L. and C. F. are supported by the European Community’s Seventh Framework Programme TriticeaeGenome (grant agreement n°FP7–212019), the Agence Nationale de la Recherche grant ANR(09-GENM-025), FranceA-griMer (201006-015-104) and the competitiveness cluster “Céréales Vallée” (http://www.cereales-vallee.org/default_gb.cfm).

References

- AGI (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- Akhunov ED, Akhunova AR, Dvorak J (2007) Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. *Mol Biol Evol* 24:539–550
- Amano N, Tanaka T, Numa H et al (2010) Efficient plant gene identification based on interspecies mapping of full-length cDNAs. *DNA Res* 17:271–279
- Astier Y, Braha O, Bayley H (2006) Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5’-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J Am Chem Soc* 128:1705–1710
- Bennett ST, Barnes C, Cox A et al (2005) Toward the \$1000 human genome. *Pharmacogenomics* 6:373–382
- Berkman P, Skarshewski A, Manoli S et al (2011a) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl Genet*: 1–10
- Berkman PJ, Skarshewski A, Lorenc MT et al (2011b) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotech J* 9:768–775
- Birney E, Stamatoyannopoulos JA, Dutta A et al (2007) Identification and analysis of functional elements in 1 % of the human genome by the ENCODE pilot project. *Nature* 447:799–816

- Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N (2012) Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* 491:705–710
- Brisson N, Gate P, Gouache D et al (2010) Why are wheat yields stagnating in Europe? A comprehensive data analysis for France. *Field Crop Res* 119:201–212
- Cantarel BL, Korf I, Robb SMC et al (2008) MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 18:188–196
- Chaisson M, Pevzner P, Tang H (2004) Fragment assembly with short reads. *Bioinformatics* 20:2067–2074
- Chantret N, Cenci A, Sabot F, Anderson O, Dubcovsky J (2004) Sequencing of the *Triticum monococcum* hardness locus reveals good microcolinearity with rice. *Mol Genet Genomics* 271:377–386
- Charles M, Belcram H, Just J et al (2008) Dynamics and differential proliferation of transposable elements during the evolution of the B and A genomes of wheat. *Genetics* 180:1071–1086
- Choulet F, Wicker T, Rustenholz C et al (2010) Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* 22:1686–1701
- Curwen V, Eyras E, Andrews TD et al (2004) The Ensembl automatic gene annotation system. *Genome Res* 14:942–950
- Devos KM, Ma J, Pontaroli AC et al (2005) Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc Natl Acad Sci U S A* 102:19243–19248
- Dolezel J, Kubalaková M, Bartos J, Macas J (2004) Flow cytogenetics and plant genome mapping. *Chromosome Res* 12:77–91
- Dolezel J, Kubalaková M, Paux E et al (2007) Chromosome-based genomics in the cereals. *Chromosome Res* 15:51–66
- Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866
- Earl D, Bradnam K, JJ St (2011) Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Res* 21:2224–2241
- Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138
- Endo TR, Gill BS (1996) The deletion stocks of common wheat. *J Hered* 87:295–307
- Erayman M, Sandhu D, Sidhu D et al (2004) Demarcating the gene-rich regions of the wheat genome. *Nucl Acids Res* 32:3546–3565
- Estill JC, Bennetzen JL (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods* 5:8
- Feldman M, Levy AA (2009) Genome evolution in allopolyploid wheat—a revolutionary reprogramming followed by gradual changes. *J Genet Genomics* 36:511–518
- Feuillet C, Eversole K (2007) Physical mapping of the wheat genome: A coordinated effort to lay the foundation for genome sequencing and develop tools for breeders. *Isr J Plant Sci* 55:307–313
- Feuillet C, Keller B (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl Acad Sci. USA* 96:8265–8270
- Feuillet C, Salse J (2009) Comparative Genomics in the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) *Plant Genetics and Genomics*. Springer, New York, pp 451–477
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trends Plant Sci* 16:77–88
- Flavell RB, Rimpau J, Smith DB (1977) Repeated sequence DNA relationship in four cereal genomes. *Chromosoma* 63:205–222
- Foley JA, Ramankutty N, Brauman KA et al (2011) Solutions for a cultivated planet. *Nature* 478:337–342

- Gill KS, Gill BS, Endo TR, Boyko EV (1996a) Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics* 143:1001–1012
- Gill KS, Gill BS, Endo TR, Taylor T (1996b) Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics* 144:1883–1891
- Gill BS, Appels R, Botha-Oberholster A-M et al (2004) A Workshop Report on Wheat Genome Sequencing: International Genome Research on Wheat Consortium. *Genetics* 168:1087–1096
- Gnerre S, MacCallum I, Przybylski D et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci. USA* 108:1513–1518
- Havlak P, Chen R, Durbin KJ et al (2004) The Atlas genome assembly system. *Genome Res* 14:721–732
- Hernandez P, Martis M, Dorado G et al (2011) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *The Plant J*: 69:377–386
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- IRGSP (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Jaffe DB, Butler J, Gnerre S et al (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13:91–96
- Janda J, Bartoš J, Šafář J et al (2004) Construction of a subgenomic BAC library specific for chromosomes 1D, 4D and 6D of hexaploid wheat. *Theor Appl Genet* 109:1337–1345
- Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, Spannagl M, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Wang J; International Wheat Genome Sequencing Consortium, Yang H, Liu X, He Z, Mao L, Wang J (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* 496:91–95
- Kubaláková M, Vrána J, Čížalíková J et al (2002) Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor Appl Genet* 104:1362–1372
- Kubaláková M, Kovářová P, Suchánková P et al (2005) Chromosome sorting in tetraploid wheat and its potential for genome analysis. *Genetics* 170:823–829
- La Rota M, Sorrells ME (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Funct Integr Genomics* 4:34–46
- Lamoureux D, Peterson DG, Li W et al (2005) The efficacy of Cot-based gene enrichment in wheat (*Triticum aestivum* L.). *Genome* 48:1120–1126
- Lander ES, Linton LM, Birren B et al (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- Lenzerini M (2002) Data integration: a theoretical perspective. Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM. Wisconsin, Madison, pp 233–246
- Leroy P, Guillhot N, Sakai H et al (2012) TriAnnot: a versatile and high performance pipeline for the automated annotation of plant genomes. *Frontiers in Plant Sciences* 3:1–14
- Li W, Gill B (2004) Genomics for cereal improvement. In: Gupta PK, Varshney RK (eds) *Cereal genomics*. Kluwer Academic Publishers, Dordrecht, pp 585–634
- Li W, Zhang P, Fellers JP et al (2004) Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J* 40:500–511
- Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967
- Liang C, Mao L, Ware D, Stein L (2009) Evidence-based gene predictions in plant genomes. *Genome Res* 19:1912–1923
- Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, Gao C, Wu H, Li Y, Cui Y, Guo X, Zheng S, Wang B, Yu K, Liang Q, Yang W, Lou X, Chen J, Feng M, Jian J, Zhang X, Luo G, Jiang Y, Liu J, Wang Z, Sha Y, Zhang B, Wu H, Tang D, Shen Q, Xue P, Zou S, Wang X, Liu X, Wang F, Yang Y, An X, Dong Z, Zhang K, Zhang X, Luo MC, Dvorak J, Tong Y, Wang J, Yang H, Li Z, Wang D, Zhang A, Wang J (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* 496:87–90

- Lobell DB, Schlenker W, Costa-Roberts J (2011) Climate Trends and Global Crop Production Since 1980. *Science* DOI:10.1126/science.1204531
- Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Massa AN, Wanjugi H, Deal KR et al (2011) Gene Space Dynamics During the Evolution of *Aegilops tauschii*, *Brachypodium distachyon*, *Oryza sativa*, and *Sorghum bicolor* Genomes. *Mol Biol Evol* 28:2537–2547
- Mayer KF, Taudien S, Martis M et al (2009) Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol* 151:496–505
- Mayer KF, Martis M, Hedley PE et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 23:1249–1263
- McFadden E, Sears E (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J Hered* 37:81–89/107
- Metzker ML (2009) Sequencing technologies – the next generation. *Nat Rev Genet* 11:31–46
- Muniz LM, Cuadrado A, Jouve N, Gonzalez JM (2001) The detection, cloning, and characterisation of WIS 2–1A retrotransposon-like sequences in *Triticum aestivum* L. and *xTriticosecale* Wittmack and an examination of their evolution in related Triticeae. *Genome* 44:979–989
- Ouyang S, Buell CR (2004) The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* 32:D360–363
- Parkhill J, Birney E, Kersey P (2010) Genomic information infrastructure after the deluge. *Genome Biol* 11:402
- Paux E, Roger D, Badaeva E et al (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J* 48:463–474
- Paux E, Sourdille P, Salse J et al (2008) A physical map of the 1-gigabase bread wheat chromosome 3B. *Science* 322:101–104
- Paux E, Sourdille P, Mackay I, Feuillet C (2011) Sequence-based marker development in wheat: Advances and applications to breeding. *Biotechnol Adv* 30:1071–1088
- Pevzner PA, Tang H, Waterman MS (2001) An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* 98:9748–9753
- Qi LL, Echalié B, Chao S et al (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 168:701–712
- Rabinowicz PD, Citek R, Budiman MA et al (2005) Differential methylation of genes and repeats in land plants. *Genome Res* 15:1431–1440
- Rounsley S, Marri P, Yu Y et al (2009) De Novo Next Generation Sequencing of Plant Genomes. *Rice* 2:35–43
- Rustenholz C, Choulet F, Laugier C et al (2011) A 3000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiol* 157:1596–1608
- Sabot F, Guyot R, Wicker T et al (2005) Updating of transposable element annotations from large wheat genomic sequences reveals diverse activities and gene associations. *Mol Genet Genomics* 274:119–130
- Šafař J, Bartoš J, Janda J et al (2004) Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J* 39:960–968
- Šafař J, Šimková H, Kubaláková M et al (2010) Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet Genome Res* 129:211–223
- Sakata K, Nagamura Y, Numa H et al (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res* 30:98–102
- Sandhu D, Gill KS (2002) Gene-Containing Regions of Wheat and the Other Grass Genomes. *Plant Physiol* 128:803–811
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci of the United States of America* 74:5463–5467

- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326:1112–1115
- Sears ER (1954) The aneuploid of common wheat. *Mo Agr Exp Sta Res Bull* 572:1–58
- Sears ER, Sears L (1978) The telocentric chromosomes of common wheat In: Ramanujams S (ed) *Proc 5th Int Wheat Genetics Symp.* Indian Agricultural Research Institute, New Delhi, India., pp 389–407
- Simkova H, Svensson JT, Condamine P et al (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* 9:294
- Simpson JT, Durbin R (2010) Efficient construction of an assembly string graph using the FM-index. *Bioinformatics* 26:i367–i373
- Simpson JT, Wong K, Jackman SD et al (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19:1117–1123
- Smith DB, Flavell RB (1975) Characterization of Wheat Genome by Renaturation Kinetics. *Chromosoma* 50:223–242
- Sorrells ME, La Rota M, Bermudez-Kandianis CE et al (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res* 13:1818–1827
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–814
- Valouev A, Ichikawa J, Tonthat T et al (2008) A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res* 18:1051–1063
- Venter JC, Adams MD, Myers EW et al (2001) The Sequence of the Human Genome. *Science* 291:1304–1351
- Vitulo N, Albiero A, Forcato C et al (2011) First survey of the wheat chromosome 5A composition through a Next Generation Sequencing approach. *PLoS ONE* 6:e26421
- Vrána J, Kubaláková M, Šimková H et al (2000) Flow sorting of mitotic chromosomes in common wheat (*Triticum aestivum* L.). *Genetics* 156:2033–2041
- Waterston RH, Lindblad-Toh K, Birney E et al (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562
- Wei F, Zhang J, Zhou S et al (2009) The physical and genetic framework of the maize B73 genome. *PLoS Genet* 5:e1000715
- Wicker T, Stein N, Albar L et al (2001) Analysis of a contiguous 211 kb sequence in diploid wheat (*Triticum monococcum* L.) reveals multiple mechanisms of genome evolution. *Plant J* 26:307–316
- Wicker T, Matthews DE, Keller B (2002) TREP: a database for Triticeae repetitive elements. *Trends Plant Sci* 7:561–562
- Wicker T, Mayer KFX, Gundlach H et al (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* 23:1706–18
- Yan L, Loukoianov A, Tranquilli G et al (2003) Positional cloning of the wheat vernalization gene VRN1. *Proc Natl Acad Sci USA* 100:6263–6268
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829
- Zhou S, Bechner M, Place M et al (2007) Validation of rice genome sequence by optical mapping. *BMC Genomics* 8:278
- Zohary D, Hopf M (2000) *Domestication of plants in the old world*, 3rd edn. Oxford University Press, Oxford

Chapter 18

Wheat Domestication: Key to Agricultural Revolutions Past and Future

Justin D. Faris

Contents

18.1	Introduction	440
18.2	The Evolution of Wheat	443
18.3	The Place and Time of Einkorn and Emmer Wheat Domestication	445
18.3.1	Domestication of Einkorn Wheat	446
18.3.2	Domestication of Emmer Wheat	446
18.4	Origin of Free-threshing Tetraploid Wheats	447
18.5	Origin of Hexaploid Wheat	448
18.6	Genetics of Domestication Loci	450
18.6.1	Brittle Rachis	451
18.6.2	Tenacious Glume	452
18.6.3	The Q Loci	454
18.6.4	The Evolution of Free-threshing Wheats	456
18.7	Wheat Evolution Under Cultivation	457
18.7.1	Capture of Genetic Variability	457
18.7.2	Further Domestication Under Cultivation	458
18.7.3	Generation of New Genetic Diversity	459
18.8	Future Needs	460
	References	460

Abstract The domestication of wheat was instrumental in the transition of human behavior from hunter-gatherers to farmers. It was a key event in the agricultural revolution that occurred about 10,000 years ago in the Fertile Crescent of the Middle East. Transitions of forms with natural seed dispersal mechanisms to forms with non-brittle rachises led to the domestication of diploid einkorn and tetraploid emmer wheat in southeast Turkey. These early domesticates were staple crops of early farmers for several thousand years before being replaced by free-threshing wheats. Allopolyploidization, mutations in genes governing threshability and other domestication related traits, and interspecific gene flow led to the formation of today's economically important bread wheat. Genetics, genomics, and archaeobotany have together

J. D. Faris (✉)

USDA-Agricultural Research Service, Cereal Crops Research Unit,
1605 Albrecht BLVD, Fargo, ND 58102-2765, USA
e-mail: justin.faris@ars.usda.gov

provided strong evidence and insights regarding the time, place, and events involved in the evolution and domestication of modern wheat, but numerous questions remain unanswered. Here, I review historical and recent findings that have shaped our current understanding of wheat domestication. Whole-genome sequence analysis, additional genetic studies, and advances in archaeology will likely address our unanswered questions in the future. A thorough and comprehensive understanding of wheat evolution and domestication will provide critical knowledge to the spawning of a new agricultural revolution, which will be necessary to provide sustenance for a rapidly increasing world population under global climate change.

Keywords Wheat · Durum · Einkorn · Emmer · Triticum · Aegilops · Evolution · Domestication · Fertile crescent · Brittle rachis · Tenacious glume · Q gene

18.1 Introduction

Before 10,000 years ago, man lived a nomadic life style as a hunter-gatherer relying on the hunting of wild animals and collecting wild plants for his food. Then, the Neolithic revolution took place where the hunter-gatherer way of life was replaced by an agrarian lifestyle. This was a crucial turning point in human history and had a profound effect on life thereafter. The Neolithic revolution took root in the Levantine Corridor and spread through the Fertile Crescent, which is located in the Middle East and encompasses a region extending from Jordan, Israel, Lebanon, and Syria through southeast Turkey and along the Tigris and Euphrates rivers through Iraq and western Iran (Fig. 18.1). This “cradle of agriculture” was the center of domestication of einkorn (*Triticum monococcum* L.) and emmer (*T. turgidum* ssp. *dicoccum* L.) wheat, which were staple crops of early civilization and close relatives of modern day wheat. These cereals were domesticated along side other important crops including barley (*Hordeum vulgare* L.), pea (*Pisum sativum* L.) Lentil (*Lens culinaris* Medikus), and chickpea (*Cicer arietinum* L.), as well as animals such as sheep (*Ovis aries*), goats (*Capra hircus*), cattle (*Bos taurus*), and pigs (*Sus scrofa*) (Zeder 2008), and they led the way for an agricultural revolution.

Nesbitt (2001) describes domestication as “. . . the process by which humans take reproductive control of plants or animals, modifying them for their own purposes.” In wheat and other cereal crops, the first and most critical modification was the acquisition of a non-brittle rachis, which limited the natural seed dispersal mechanisms of the wild forms and allowed early farmers to harvest the grain much more efficiently without spikelets dropping to the ground prematurely and being lost. Other modifications included larger seeds, loss of seed dormancy, the free-threshing character, enhanced grain quality, and others (Harlan et al. 1973). These changes resulted in domesticated forms that relied on farmers for their propagation and also allowed mechanized cultivation on a large scale.

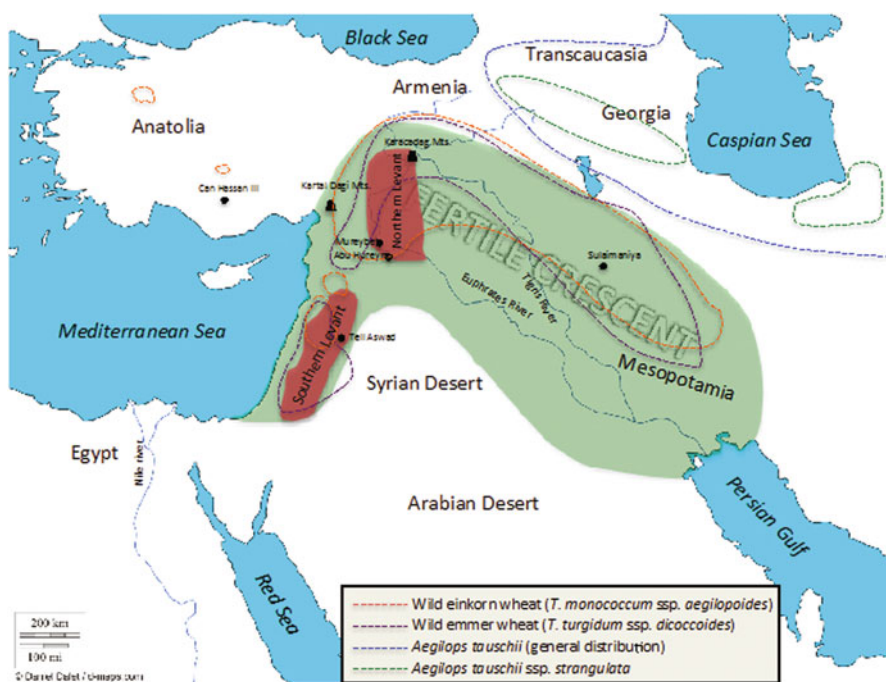


Fig. 18.1 Map of the ancient Middle East showing the Fertile Crescent (green). The Southern and Northern Levant regions are indicated by the brown shaded areas. Archaeological sites mentioned in the text are indicated by black circles (villages) and trapezoids (mountain ranges). The distributions of wild wheats are indicated by the dotted lines (see legend at the bottom of the figure). The basic map was obtained from d-maps.com. (http://d-maps.com/carte.php?lib=fertile_crescent_map&num_car=5852&lang=en)

Today, about 430 million tonnes of the fully domesticated free-threshing hexaploid and tetraploid wheats known as common, or bread, wheat (*T. aestivum* ssp. *aestivum* L.) and durum, or macaroni, wheat (*T. turgidum* ssp. *durum* L.), respectively, are produced annually and provide about a fifth of the calories consumed by humans worldwide (<http://faostat.fao.org>). Bread wheat accounts for about 95 % of the total wheat crop and is used to make bread, cookies, cakes, crackers, pastries, and noodles, whereas durum wheat accounts for the remaining 5 % and is used to make pasta and other semolina products. Due to the rate of the world's population growth, the demand for wheat is expected to increase by 40 % by 2030 (Dixon et al. 2009). In order to meet this demand, an annual increase in yield of 2 % is needed and the amount of agricultural land needs to be stabilized. These gains will need to come by way of genetic improvements and enhanced understanding of plant biology. Advancing our knowledge and understanding of wheat evolution and the genetic mechanisms underpinning the core domestication events that shaped today's wheat plant may provide new clues as to how the genetic diversity available in the wild wheat progenitors and relatives can be tapped into and exploited to initiate a modern agricultural revolution under a changing global climate.

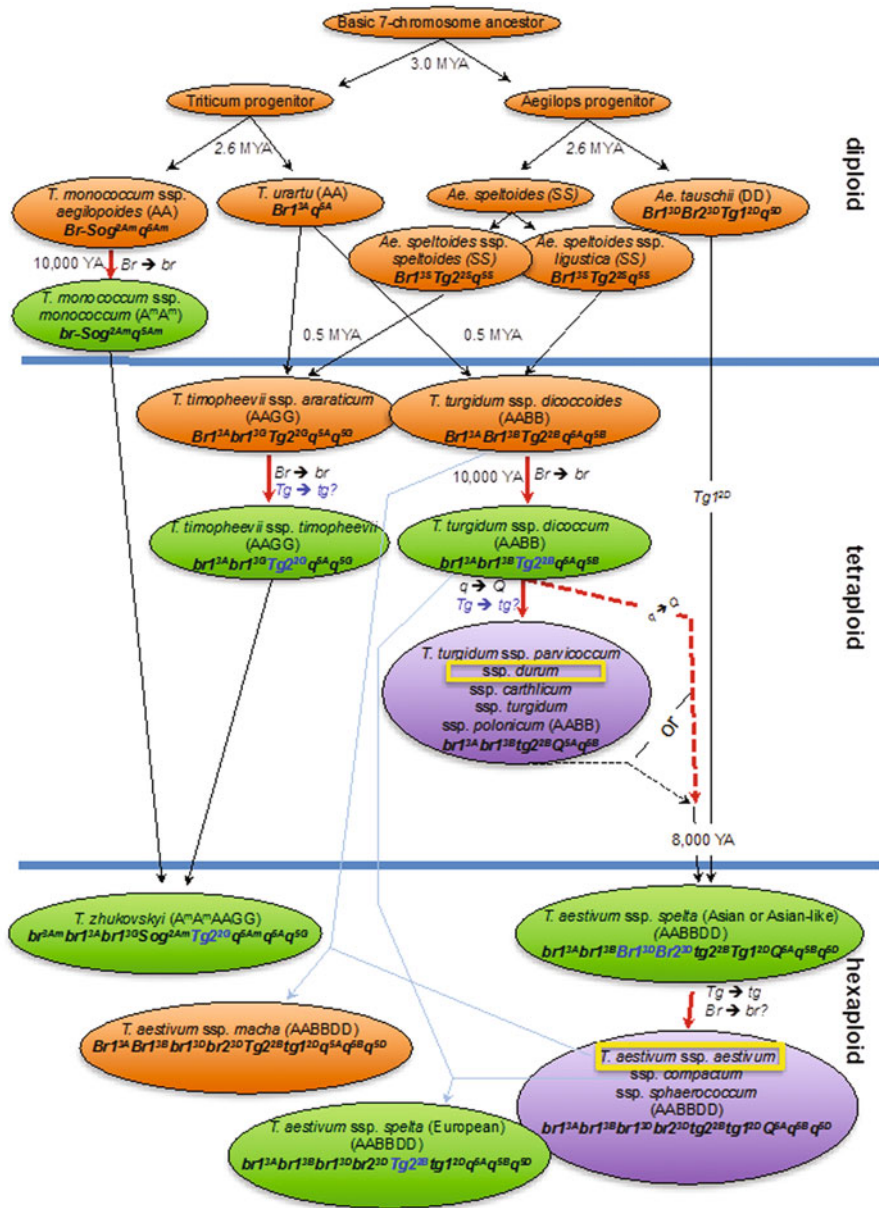


Fig. 18.2 The evolutionary lineages involving *Triticum* wheat species. The diploid, tetraploid, and hexaploid species are separated by blue bars. Orange, green, and purple colors indicate species with brittle rachis and hulled seed, species with non-brittle rachis and hulled seed, and species with a fully tough rachis and free-threshing seed, respectively. Red arrows indicate occurrences of transitions involving one or more of the major domestication genes *Br*, *Tg*, or *Q*. Genotypes of major domestication genes are indicated in bold below the taxonomical names and their genome

18.2 The Evolution of Wheat

It was determined nearly a century ago that the cultivated wheat species of the genus *Triticum* have chromosome numbers of $2n = 14$, 28, and 42. This indicated that the basic Triticeae genome was organized into seven chromosomes ($1x = 7$) and the various *Triticum* species consisted of diploids ($2n = 2x = 14$), tetraploids ($2n = 4x = 28$), and hexaploids ($2n = 6x = 42$) (Sax 1922; Kimber and Sears 1987).

The diploid progenitors and close relatives of modern wheat radiated from a common ancestor about 3 million years ago (MYA) and gave rise to the *Triticum* and *Aegilops* taxa (Fig. 18.2). The *Triticum* group consisted of the A-genome diploids *T. urartu* Tumanian ex Gandylan ($2n = 2x = 14$, AA (the capital letters represent the genome constitution)) and *T. monococcum* ssp. *aegilopoides* (Link) Thell. ($2n = 2x = 14$, AA). Johnson and Dhaliwal (1976) determined that they are valid biological species. Also evolving from the common seven-chromosome ancestor were numerous diploid *Aegilops* species including *Ae. tauschii* Coss. ($2n = 2x = 14$, DD) and a progenitor to the *Aegilops* Sitopsis section, which gave rise to the S-genome containing *Aegilops* species including *Ae. speltoides* Tausch ($2n = 2x = 14$, SS).

The only domesticated diploid wheat is einkorn (*T. monococcum* ssp. *monococcum* L., $2n = 2x = 14$, A^mA^m), which was domesticated from ssp. *aegilopoides* through the acquisition of a non-brittle rachis (Fig. 18.2). The evolution and formation of the cultivated forms of polyploid wheat followed two basic lineages, both of which involved two amphiploidization events. These events resulted from the hybridization of two different species followed by spontaneous chromosome doubling of the F₁ hybrid through the functioning of meiotic restitution division (non-reduced) gametes. One lineage began with hybridization of *T. urartu* (Dvorak et al. 1993) and *Ae. speltoides*, or a close relative thereof (Sarkar and Stebbins 1956; Riley et al. 1958), which led to the formation of the wild emmer wheat *T. timopheevii* ssp. *araraticum* Jakubz. ($2n = 4x = 28$, AAGG) containing a pair of A genomes from *T. urartu* and a pair of G genomes, which are considered to be a divergent form of the S genome of the *Aegilops* progenitor (Rodriquez et al. 2000). *T. timopheevii* ssp. *araraticum* has a brittle rachis conferred by the *Br1^{3A}* gene. A mutation in *Br1^{3A}* led to a non-brittle rachis and the domestication of this form to *T. timopheevii* ssp. *timopheevii* (Zhuk.) Zhuk ($2n = 4x = 28$, AAGG). *T. timopheevii* was never cultivated as a significant crop and grows only in a limited region of Georgia. Therefore, it was probably a secondary domesticate (Nesbitt and Samuel 1996).

The hexaploid wheat belonging to this lineage is *T. zhukovskiyi* Menabde et Ericzjan ($2n = 6x = 42$, A^mA^mAAGG), which resulted from a hybridization between *T. timopheevii* ssp. *timopheevii* and domesticated einkorn wheat (Jakubziner 1958; Johnson 1968). Like ssp. *timopheevii*, *T. zhukovskiyi* is not cultivated and tends to

Fig. 18.2 constitutions. Homozygosity is inferred at each locus, and genotypes are indicated only once to save space. Genotypes in *blue* are suggested but no experimental evidence is available. *Blue arrows* represent events that occurred to give rise to hexaploid subspecies that formed subsequent to *T. aestivum* ssp. *aestivum*. Durum and common wheat, the two modern widely cultivated forms of polyploid wheat, are highlighted with *yellow rectangles*

be found in Western Georgia as an admixture with *T. timopheevii* and einkorn wheat (Nesbitt and Samuel 1996). Although this constitutes an interesting evolutionary lineage of wheat, it did not result in the formation of any of today's economically important wheats. Therefore, little attention will be devoted to the species of this lineage in the remainder of this review.

Like the first lineage, the second also began with a hybridization event between *T. urartu* (Dvorak et al. 1993) and a close relative of *Ae. speltooides* (Dvorak and Zhang 1990; Blake et al. 1999; Huang et al. 2002; Chalupska et al. 2008; Salse et al. 2008) but one of a different subspecies than the one involved in the first lineage (Kilian et al. 2007; Fig. 18.2). This event led to the formation of the tetraploid wild emmer wheat *T. turgidum* ssp. *dicoccoides* (Körn.) Thell ($2n = 4x = 28$, AABB genomes). Although both wild emmer wheat species (*T. turgidum* spp. *araraticum* and *dicoccoides*) obtained a pair of genomes from the S genome-containing *Ae. speltooides*, significant divergence has since occurred such that the B, G, and S genomes, while still related, are quite distinct (Zhang et al. 2002; Kilian et al. 2007). Like ssp. *araraticum*, ssp. *dicoccoides* has a brittle rachis, and mutations in *Br* loci led to the domesticated emmer subspecies *T. turgidum* ssp. *dicoccum* (Schrank) Schübl ($2n = 4x = 28$, AABB genomes), which has a non-brittle rachis and hulled seed (Fig. 18.3). The second amphiploidization event of this lineage resulted when the diploid goat grass *Ae. tauschii* hybridized with a *T. turgidum* subspecies (Kihara 1944; McFadden and Sears 1946; Fig. 18.2). The resulting hexaploid was hulled due to the presence of the *Tgl* gene from *Ae. tauschii* (McFadden and Sears 1946). This subspecies may have been similar to Asian spelta (*T. aestivum* ssp. *spelta* L., $2n = 6x = 42$, AABBDD) (Figs. 18.2, 18.3). Evolution of this species through the acquisition of the free-threshing character (see below) resulted in the free-threshing hexaploid bread wheat *T. aestivum* ssp. *aestivum* L., $2n = 6x = 42$, AABBDD) (Figs. 18.2, 18.3), one of the most economically important crops in the world today.

Other hexaploid *T. aestivum* spp. include *compactum* (Host) MacKey, *sphaerococcum* (Percival) MacKey, *macha* Dekapr. et Menabde, and European *spelta* L. Subspecies *compactum* (club wheat) and *sphaerococcum* (shot wheat) are both free-threshing and differ from *aestivum* by single genes that arose through mutation. Ssp. *compactum* carries the *C* gene, which confers a compact spike and *sphaerococcum* carries the *S* gene for spherical grains. Ssp. *compactum* is grown today in a few isolated areas of Europe, the Near East, and the northwestern U.S., whereas *sphaerococcum* is confined mostly to India. Spp. *macha* and European *spelta* are not free-threshing and resemble primitive forms, but are not progenitors to *aestivum*. Ssp. *macha* likely arose through a more recent hybridization between *aestivum* and the emmer wheat *T. turgidum* ssp. *dicoccum* (Dvorak and Luo 2001), and there is now much evidence demonstrating that European *spelta* formed from a cross between *T. aestivum* ssp. *compactum* and domesticated emmer (*T. turgidum* ssp. *dicoccum*) (Bertsch 1943; MacKey 1966; Blatter et al. 2002, 2004; Yan et al. 2003; Fig. 18.2).

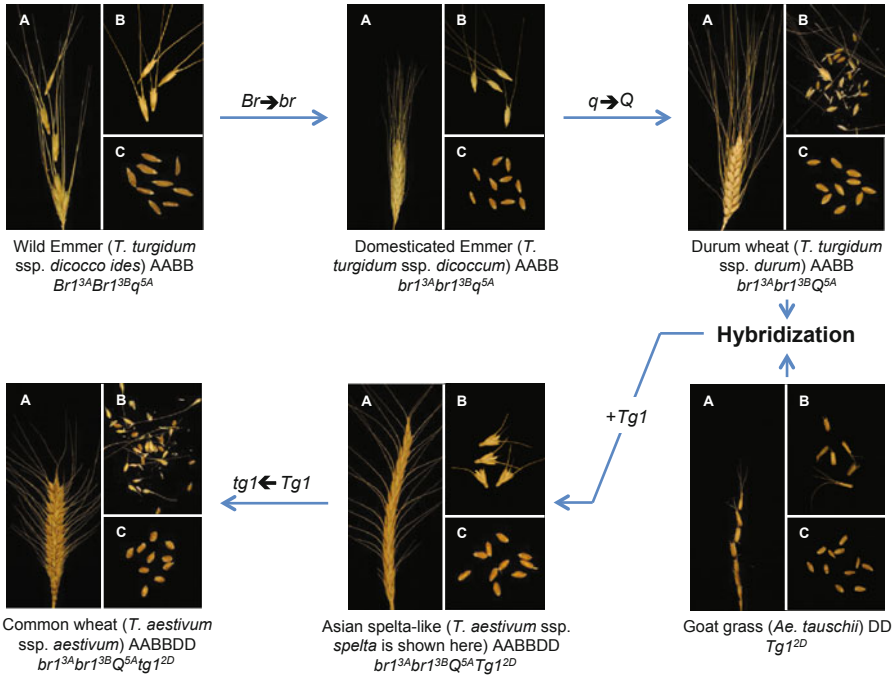


Fig. 18.3 The phenotypes of wheat species involved in the evolution of modern cultivated wheat and major transitions at the primary domestication genes ($Br1^{3A}$, $3B$, q^{5A} , and $Tg1^{2D}$). The spike, result of moderate hand threshing, and seed are shown for each in panels A, B, and C, respectively. Although direct evidence is lacking, the most likely scenario in which a free-threshing (Q -containing) tetraploid hybridized with the D genome diploid *Ae. tauschii* to produce hexaploid wheat is shown. Also, Asian spelta is shown to represent the primitive progenitor to free-threshing common wheat because it is likely that the progenitor was very similar. The common names, taxonomical names, genome constitutions, and genotypes are indicated below each panel

18.3 The Place and Time of Einkorn and Emmer Wheat Domestication

Today, findings from archaeological digs combined with molecular genetics experiments have provided many answers regarding wheat domestication. Agriculture originated in the Fertile Crescent approximately 10,000 years ago, but the earliest gathering of wild emmer wheat has been dated to 19,000 years before present (BP) in the Middle East indicating that humans collected wild grains for some 10,000 years before domestication took place. Furthermore, Tanno and Willcox (2006) argue that cereals were actually cultivated for over a thousand years before the emergence of domesticates in what would be considered the first phase of cultivation. In the second phase of cultivation, domesticated forms consisting of einkorn and emmer wheat were grown by early farmers. These wheats acquired a non-brittle rachis, which allowed early farmers to efficiently harvest the grain without the spikes shattering and

falling to the ground before harvest. Later, further domestication occurred through the acquisition of the free-threshing character. This led to complete replacement of the hulled einkorn and emmer wheats with the free-threshing durum and bread wheat, although replacement was not rapid.

18.3.1 Domestication of Einkorn Wheat

Wild (*T. monococcum* ssp. *aegilopoides*) and domesticated einkorn wheat (*T. monococcum* ssp. *monococcum*) are very similar morphologically except that the former has a brittle rachis and the latter a non-brittle rachis. The distribution of wild einkorn in the Middle East is rather widespread and is found from the Balkans to Iran primarily in the northern and eastern parts of the Fertile Crescent growing as a weed along roadsides and fields (Nesbitt and Samuel 1996; Fig. 18.1). Man collected grains of wild einkorn for some time before cultivating it, because brittle-rachis einkorn wheat has been found, for example, in the prehistoric settlement of Mureybit of the northern Levantine Corridor dated about 10,000 BP (Renfrew 1973; Fig. 18.1). Then, the mutation that resulted in a non-brittle rachis led to the domestication of einkorn wheat. Using amplified fragment length polymorphism (AFLP) DNA fingerprinting, Heun et al. (1997) located the site of einkorn domestication to the Karacadag region of southeastern Turkey in the northern Levantine Corridor. From here, wild and domesticated einkorn were grown side by side throughout the Fertile Crescent with the non-brittle rachis type gradually replacing wild einkorn. Cultivation of domesticated einkorn then spread to other parts of the Middle East and southern Europe, and eventually central and western Europe (Feldman 2001). Today, einkorn wheat is a relic crop that is grown somewhat in parts of Turkey, Italy, and the former Yugoslavia primarily for animal feed.

18.3.2 Domestication of Emmer Wheat

Wild emmer (*T. turgidum* ssp. *dicoccoides*) is the only true wild polyploid wheat of the lineage, and it is the progenitor of today's durum and bread wheat cultivars. Therefore, the discovery of wild emmer and documentation of its distribution (Aaronsohn 1910) contributed substantially to our understanding of the events that led to the domestication of modern wheats. Unlike wild einkorn wheat, wild emmer does not grow as a weed but only as a truly wild plant mostly confined to relatively undisturbed habitats. Therefore, wild einkorn grows over a much wider area than wild emmer, and the latter has not spread much outside of the Fertile Crescent growing in a region extending from the southern Levant across Israel and Lebanon to southeastern Turkey and across northern Iraq and northwestern Iran (Fig. 18.1). However, the region of wild emmer habitat is not continuous and can be divided into northern and southern subpopulations.

As with einkorn wheat, wild emmer wheat was cultivated for some time before mutants with a non-brittle rachis appeared. Then, wild emmer was likely cultivated for several hundred more years as a mixture with the non-brittle rachis form known as domesticated emmer (*T. turgidum* ssp. *dicoccum*) (Kislev 1984). The archaeological record indicates that, as with domesticated einkorn wheat, domesticated emmer first appeared in the southern Levant and in southeastern Turkey about 9,500–9,000 BP (Nesbitt and Samuel 1996), but the question of where emmer was first domesticated has been debatable. The fact that domesticated emmer showed up in both the northern and southern Levant almost simultaneously suggests that emmer was domesticated in either the northern or southern part and then rapidly spread to the other. Using AFLP analysis, Ozkan et al. (2002) showed that domesticated emmer was more closely related to the northern wild emmer populations than to the southern populations. On the contrary, Mori et al. (2003) evaluated chloroplast microsatellite variation in a more complete set of wild emmer accessions and concluded that emmer was domesticated in the Kartal Dagi mountains of southeastern Turkey, but also suggested emmer may have been domesticated a second time elsewhere. Using the accessions of Mori et al. (2003), Ozkan et al. (2005) conducted AFLP analysis and showed that the Kartal Dagi region was not the site of emmer domestication, and instead suggested that emmer was domesticated in the Karacadag region, the Sulaimaniya region, or both regions independently. Luo et al. (2007) used RFLP fingerprinting to show that emmer was unlikely domesticated in the Sulaimaniya region, but was likely domesticated in the Karacadag region. Substantial gene flow between the northern domesticated emmer population and the southern wild emmer population resulted in high levels of diversity in the southern Levant and led to the development of northern and southern subpopulations of domesticated emmer (Luo et al. 2007). Therefore, both einkorn and emmer were most likely of monophyletic origin and both were domesticated in essentially the same place.

Free-threshing derivatives of domesticated emmer, such as the extinct tetraploid *T. turgidum* ssp. *parvicoccum*, appear in the archaeological record shortly after domesticated emmer (Kislev 1980). In spite of being non-free-threshing, domesticated emmer was the most abundant wheat crop in the Middle East during the Prepottery Neolithic B period, and continued to be a major crop for several thousand years. About 8,000 BP it spread from the northern Fertile Crescent, presumably along with *T. turgidum* ssp. *parvicoccum*, south to Mesopotamia and west to Anatolia, the Mediterranean basin, and Europe (Feldman 2001). It arrived in Egypt, central Asia, and India about 6,000 BP and was the dominant cereal crop in all these regions up until about 3,000 BP when it was largely replaced by free-threshing durum wheat. Today, domesticated emmer is a relic crop grown only in limited areas of the Middle East and south Asia.

18.4 Origin of Free-threshing Tetraploid Wheats

The archaeological record indicates that free-threshing tetraploid wheats appeared about 8,000–9,000 BP in the Prepottery Neolithic B period, about the same time as domesticated emmer. These early finds occurred in Tell Aswad and other Syrian

sites as well as Can Hassan III in southern Turkey (Kislev 1980). This free-threshing tetraploid with very small grains and compact spikes was considered to be an extinct subspecies and given the name *T. turgidum* ssp. *parvicoccum*. It was assumed that this wheat was derived from domesticated emmer (*T. turgidum* ssp. *dicoccum*) and that domesticated emmer was grown for some time before ssp. *parvicoccum* appeared (Feldman 2001). Ssp. *parvicoccum* may have been grown as an admixture with domesticated emmer wheat, or it may have been grown as a separate crop. In either case, it spread along with domesticated emmer throughout the Fertile Crescent and was grown for several millennia in the Middle East. However, some have questioned the existence of ssp. *parvicoccum* due to a limited number of completely characterized samples (Nesbitt 2001).

Durum wheat (*T. turgidum* ssp. *durum*) evolved from domesticated emmer wheat possibly by way of ssp. *parvicoccum*. The first durum wheat was found at Can Hassan III and dated about 6,500–7,500 years BP (Hillman 1978), but durum wheat was not established as a prominent crop until about 2,300 years BP (Feldman 2001). Today, durum is a major crop well adapted to dry climates and used for macaroni and semolina products. It is primarily grown in the Great Plains region of the U.S. and Canada, Russia, India, Italy, and the Middle East.

Most other tetraploid wheat subspecies are free-threshing and probably arose relatively recently. These include spp. *turgidum*, *turanicum*, *polonicum*, *carthlicum*, and others all of which are quite similar to ssp. *durum* and differ by only a few traits. Some, for example ssp. *carthlicum*, probably arose by hybridization between the free-threshing hexaploid *T. aestivum* and another tetraploid (Kuckuck 1979).

18.5 Origin of Hexaploid Wheat

Both free-threshing and non free-threshing forms of cultivated hexaploid wheat exist, but wild progenitors of cultivated hexaploid wheat do not. Hexaploid wheat originated as a result of hybridization between an AB genome-containing tetraploid and the diploid goat grass *Ae. tauschii*, which contributed the D genome (Kihara 1944; McFadden and Sears 1946). A number of studies have pointed to *Ae. tauschii* ssp. *strangulata* as being the donor of the D genome as opposed to ssp. *tauschii* (Nishikawa 1974; Nishikawa et al. 1980; Jaaska 1978, 1980, 1981; Dvorak et al. 1998). *Ae. tauschii* ssp. *strangulata* is distributed in two regions: Transcaucasia and an area of Iran southeast of the Caspian Sea (Fig. 18.1). Therefore, it was thought that the hybridization event that formed hexaploid wheat must have occurred in one of these two areas and several lines of research involving the evaluation of collections of *Ae. tauschii* populations pointed to the area of Iran southeast of the Caspian Sea as the most probable birthplace of hexaploid wheat (Jaaska 1980; Nakai 1979; Dvorak et al. 1998), which most likely involved multiple amphiploidization events between *T. turgidum* and *Ae. tauschii* with subsequent intercrossing that led to the formation of a single gene pool (Dvorak et al. 1998; Lelley et al. 2000).

Because *T. turgidum* was largely confined to the Fertile Crescent and the distribution of *Ae. tauschii* primarily occupies northern Iran, Transcaucasia, and Afghanistan, the hybridization event(s) that resulted in hexaploid wheat probably did not occur until expansion of domesticated emmer subpopulations in the northern Levant overlapped with the primary distribution areas of *Ae. tauschii* in Transcaucasia and south of the Caspian Sea thereby providing the birthplace of hexaploid wheat. However, the South Caspian being the birthplace of hexaploid wheat does not agree with the archaeological record (Nesbitt and Samuel 1996), because the earliest records of hexaploid wheat date to 8,800 to 8,400 BP identified from several areas including Can Hassan III in southern Turkey and Abu Hureyra in Syria (Hillman 1978; Moore et al. 2000; de Moulins 2000; Fairbairn et al. 2002; Fig. 18.1). In line with this, Giles and Brown (2006) reported finding ancient *Ae. tauschii* populations in Syria and Turkey, and obtained results suggesting that the first hybridization event between *T. turgidum* and *Ae. tauschii* that gave rise to hexaploid wheat could have occurred in southeastern Turkey or northern Syria, within the Fertile Crescent near the first archaeological findings. Therefore, while most studies agree that hexaploid wheat is of polyphyletic origin involving more than one *Ae. tauschii*, the exact site of the origin of hexaploid wheat is yet uncertain.

The *T. turgidum* parent involved in the formation of hexaploid wheat is also yet a matter of debate. It is generally accepted that ssp. *dicoccoides* was not the AB-genome donor because, if it were, the resulting hexaploid would have a brittle rachis, and therefore little chance of being selected by farmers (Kimber and Sears 1987). The most probable tetraploid progenitors to hexaploid wheat are domesticated emmer (ssp. *dicoccum*) or an extinct free-threshing subspecies such as ssp. *parvicoccum*, which appear about the same time in the archaeological record. It is interesting to note that Kerber (1964) extracted the AB genome components from hexaploid wheat, and the resulting AB-tetraploids were very similar in spike morphology to ssp. *parvicoccum*. However, some genetic evidence based on genes for waxiness points to domesticated emmer as the AB progenitor (Tsunewaki 1966), while other research based on meiotic restitution suggests free-threshing durum wheat (*T. turgidum* ssp. *durum*) could have been involved (Matsuoka and Nasuda 2004).

Regardless of the subspecies involved, it is certain that *T. turgidum* was the donor of the AB genomes and *Ae. tauschii* donated the D genome to hexaploid wheat (McFadden and Sears 1946). The first hexaploid had hulled seed due to the *Tg1* (tenacious glume; see below) gene acquired from *Ae. tauschii*, and therefore would have been very similar to *T. aestivum* ssp. *spelta* (McFadden and Sears 1946; Kerber and Rowland 1974). Today, two forms of *T. aestivum* ssp. *spelta* exist and are classified as European and Asian spelta. European spelta first appeared in Europe in the Early Bronze Age (4200–3500 BP) near the Swiss lake district and elsewhere (Nesbitt 2001) several thousand years after the appearance of free-threshing hexaploid wheat. It was once thought that European spelta could have been the progenitor to free-threshing hexaploid wheat, but it is now known that it arose more recently as a result of hybridization between *T. aestivum* ssp. *compactum* and domesticated emmer (*T. turgidum* ssp. *dicoccum*) (Bertsch 1943; MacKey 1966; Blatter et al. 2002, 2004; Yan et al. 2003).

Kuckuck (1959) found a spelta-like form of wheat growing in north Iran, and other populations of this Asian spelta are now known to grow in Afghanistan, Tadshikistan, Armenia, and other areas of the Middle East region. Genetic studies have indicated that Asian and European spelta have separate origins (MacKey 1966; Liu and Tsunewaki 1991; Luo et al. 2000), leaving open the possibility that Asian spelta could be the primitive form of common wheat. However, there is no reliable archaeological evidence for the existence of spelta in the regions where hexaploid wheat is thought to have originated, and it is fairly certain that cultivation of free-threshing common wheat preceded the cultivation of hulled hexaploid wheat making it unlikely that Asian spelta is ancestral to common wheat and more likely that it, like European spelta, is a secondary derivative of common wheat.

Neither genetic nor archaeological evidence provide answers as to when and where either type of spelta originated, but it is clear that the first hexaploid must have been spelta-like (McFadden and Sears 1946; Kerber and Rowland 1974). The archaeological record would suggest that neither is ancestral to common wheat, and genetic analysis further rules out European spelta as a primitive form. Therefore, the first hulled hexaploid may have been short-lived and now extinct. This would suggest that the transition from hulled to free-threshing wheat occurred very rapidly, and was probably the result of a mutation of *Tg1* to *tg1* (see below).

The earliest findings of free-threshing ssp. *aestivum* are from Can Hassan III about 8,500 BP, which agrees fairly well with the genetic evidence for the origin of hexaploid wheat occurring about 8,000 BP (Huang et al. 2002). Later finds were unearthed in western Iran, northern Iraq, and western Anatolia followed by finds in the Mediterranean basin and Mesopotamia. Free-threshing common wheat then spread from these areas about 6,000 BP to the Nile Basin, central and western Europe, and Asia.

18.6 Genetics of Domestication Loci

Transitions in three major genes during wheat evolution ultimately yielded free-threshing fully domesticated bread wheat. Those three major genes are *Br*, *Tg*, and *q*, which, in their primitive form, confer a brittle rachis, tenacious glume, and the non free-threshing character, respectively. Mutations in the *Br1* loci on chromosomes 3A and 3B, *Tg1* on 2D, and *q* on 5A had the most profound impacts on domestication characters (Fig. 18.3), but homoeologous copies of each of these genes, and in some cases alternate genetic loci, have also been shown to govern and/or influence domestication traits. Therefore, a relatively detailed synopsis of our understanding of these three major genes, their homoeologs, and other relevant loci is provided below.

18.6.1 Brittle Rachis

A mechanism of natural seed dispersal is a hallmark trait of a wild plant species because it is essential to ensure the spread and propagation of the species. As essential as it is to wild plants, it is as detrimental to cultivated plant species because the fruits or seeds fall to the ground at maturity and are lost. Therefore, the loss of the ability to naturally disperse seed, i.e. the change from a brittle rachis to a non-brittle rachis, was one of the first and most essential domestication traits acquired by the cultivated wheat forms. Seed dispersal systems depend on the formation of abscission zones at particular sites that allow breakage and subsequent dispersal of fruits or seeds. The two basic types of disarticulation found in wheat are spike type, where breakage occurs at the base of the spike and the whole spike is dispersed as a single unit, and spikelet-type, which is further classified into either barrel- (B) or wedge-shaped (W) disarticulation, depending on the disarticulation products. Abscission at the bottom of the spikelet base leaving an adjacent rachis fragment attached behind the spikelet is considered B-type disarticulation. W-type disarticulation is when abscission occurs such that a rachis fragment is left attached below each spikelet.

Wild einkorn wheat (*T. monococcum* ssp. *aegilopoides*) undergoes W-type disarticulation whereas domesticated einkorn, *T. monococcum* ssp. *monococcum* has a non-brittle rachis. Sharma and Waines (1980) showed that non-brittleness in ssp. *monococcum* was controlled by two complementary recessive genes. To my knowledge, the chromosomal locations of these genes have not been determined and therefore their relationships with the other more characterized brittle rachis genes in wheat (see below) are unknown. While the rachis of ssp. *monococcum* is not as fragile as that of ssp. *aegilopoides*, it is not very tough, and moderate pressure causes breakage of the rachis leading to spikelet segments resembling those of wild einkorn. Consequently, the threshing of both forms leads to hulled grains in the form of spikelets with W-type disarticulation. Therefore, the domestication of einkorn wheat may be considered incomplete, or partial, because it involved only one or a few of the several important steps needed toward complete domestication.

Similar transitions occurred in the domestication of cultivated emmer (*T. turgidum* ssp. *dicoccum*) from wild emmer (*T. turgidum* ssp. *dicoccoides*). Studies using genetic stocks where individual pairs of wild emmer chromosomes were substituted for homologous pairs of durum chromosomes have shown that wild emmer chromosomes 3A and 3B harbor genes conferring the brittle rachis trait (Watanabe and Ikebata 2000). Molecular mapping analysis using recombinant inbred chromosome lines derived from the same stocks evaluated by Watanabe and Ikebata (2000) indicated that the *Br* genes are located on the short arms of chromosomes 3A (*Br1^{3A}*) and 3B (*Br1^{3B}*) and that they are likely homoeologous (Nalam et al. 2006; Table 18.1). Both genes lead to W-type disarticulation (Fig. 18.3). This and other studies have also indicated that the *Br1* loci in tetraploid wheat are homoeologous with the *Btr1/Btr2* loci, which confer a brittle rachis in barley (Nalam et al. 2006; Li and Gill 2006). These studies indicate that *Br1^{3A}* was derived from *T. urartu* and *Br1^{3B}* from the B-genome donor, and that mutations at both loci were needed to confer the non-brittle

rachis of domesticated emmer. However, like domesticated einkorn wheat, domesticated emmer does not have a very tough rachis and disarticulation is similar to wild emmer when sufficient pressure is applied. Therefore, threshing leads to hulled seed in the form of spikelets for both wild and domesticated emmer (Fig. 18.3).

At least two studies regarding the genetics and mapping of rachis brittleness in wild emmer have reported a *Br* gene on the long arm of chromosome 2A (Peng et al. 2003; Peleg et al. 2011; Table 18.1). While neither study reported the type of disarticulation conferred by the 2A locus, this would suggest that the genetic system involved in controlling spikelet disarticulation is under complex regulation or perhaps the 2A locus represents an independent genetic pathway.

The mutations in the wild emmer *Br* loci happened before the amphiploidization event that gave rise to hexaploid wheat. Therefore, with the exception of *T. aestivum* ssp. *macha* (which was likely formed secondarily) none of the hexaploid subspecies have the primitive *Br1^{3A}* or *Br1^{3B}* alleles and therefore do not have brittle rachises conferred by these genes. A hexaploid semi-wild wheat landrace was found in Tibet and reported to have a fragile rachis and W-type disarticulation (Cao et al. 1997). It was later determined that rachis brittleness in this line was due to a *Br* gene (*Br1^{3D}*) on the short arm of 3D, which was likely derived from *Ae. tauschii* and homoeologous to *Br1^{3A}* and *Br1^{3B}* (Chen et al. 1998; Table 18.1). However, *Ae. tauschii* has B-type disarticulation and so it was suggested that *Br1^{3D}* likely confers B-type disarticulation in *Ae. tauschii* but W-type in a hexaploid background. Further work by Li and Gill (2006) indicated that B-type disarticulation in *Ae. tauschii* is conferred by a *Br* gene (*Br2^{3D}*) on the long arm of chromosome 3D. They concluded that B- and W-type disarticulations are controlled by different genes with *Br1^{3A}*, *Br1^{3B}*, and *Br1^{3D}* controlling W-type and *Br2^{3D}* controlling B-type. In either case, the *Br* gene(s) present in the *Ae. tauschii* progenitor must have undergone mutation very soon after the hybridization event that gave rise to hexaploid wheat, or if the *Ae. tauschii* progenitor carried only *Br2^{3D}*, its effects are greatly diminished in a hexaploid background.

18.6.2 Tenacious Glume

The primitive non-domesticated wheat forms had tough glumes that tightly enveloped the seed in order to protect it during natural seed dispersal. Domesticated einkorn wheat, the only cultivated diploid wheat, is not free-threshing because it has tough adherent glumes that do not allow the seed to be easily separated from the spikelets. Investigation of a spontaneous free-threshing mutant of domesticated einkorn wheat, referred to as *T. sinskajae*, indicated that a single recessive gene designated *sog* controlled the soft glume trait and mapped to the short arm of chromosome 2A^m (Taenzler et al. 2002; Sood et al. 2009).

The glumes of wild and domesticated emmer are tough, hold the kernels tightly, and prohibit the free-threshing trait (Fig. 18.3). Simonetti et al. (1999) evaluated a tetraploid mapping population derived from a cross between *T. turgidum* spp.

Table 18.1 The three principal traits affected by mutation leading to wheat domestication and their associated genetic loci

Trait	Gene	Chrom. arm	Species that acquired mutation to domestic form (ploidy)	Reference
Brittle rachis	<i>Br1^{3A}</i>	3AS	<i>T. turgidum</i> ssp. <i>dicoccum</i> (4 ×)	Watanabe and Ikebata (2000) Nalam et al. (2006) Li und Gill (2006)
	<i>Br1^{3B}</i>	3BS	<i>T. turgidum</i> ssp. <i>dicoccum</i> (4 ×)	Watanabe and Ikebata (2000) Nalam et al. (2006) Li und Gill (2006)
	<i>Br1^{3D}</i>	3DS	<i>T. aestivum</i> ssp. <i>aestivum</i> (6 ×)	Chen et al. (1998)
	<i>Br2^{3D}</i>	3DL	<i>T. aestivum</i> ssp. <i>aestivum</i> (6 ×)	Li und Gill (2006)
	<i>Br4^{2A}</i>	2AL	<i>T. turgidum</i> ssp. <i>dicoccum</i> (4 ×)	Peng et al. (2003, 2011)
Tenacious glume	<i>Tg1^{2D}</i>	2DS	<i>T. aestivum</i> ssp. <i>aestivum</i> (6 ×)	Jantasuriyarat et al. (2004) Nalam et al. (2007) Sood et al. (2009)
	<i>Tg2^{2B}</i>	2BS	<i>T. turgidum</i> ssp. <i>parvicoccum</i> (4 ×) ^a	Simonetti et al. (1999)
Free-threshing	<i>Q^{5A}</i>	5AL	<i>T. turgidum</i> ssp. <i>parvicoccum</i> (4 ×) ^a or <i>T. aestivum</i> ssp. <i>aestivum</i> (6 ×)	Faris et al. (2005) Simons et al. (2006)

^a This subspecies is not known for certain and could have been *parvicoccum*, *durum*, or perhaps another tetraploid subspecies

dicoccoides and *durum* for quantitative trait loci (QTLs) associated with the free-threshing trait. One QTL corresponded to the free-threshing locus *Q* on the long arm of chromosome 5A (see below) and another with major effects mapped to the short arm of chromosome 2B. The latter QTL was located in a position apparently syntenic with the tough glume gene *Tg1^{2D}* on chromosome 2D (see below). It is possible that the gene underlying the 2B QTL (*Tg2^{2B}*) and *Tg1^{2D}* (Table 18.1) are homoeologous, or that *Tg2^{2B}* is homoeologous with *Sog* on 2A^m in einkorn wheat, but appropriate comparative mapping experiments to address these matters have not been conducted. Another matter to be addressed is whether or not domesticated emmer (*T. turgidum* ssp. *dicoccum*) possesses *Tg2^{2B}*, or if the tough glume, non-free-threshing trait of domesticated emmer is due to the fact it carries the *q* allele on 5A. The finding that *T. turgidum* ssp. *dicoccum* var. *liguliforme*—a variety of domesticated emmer with a dense spike—carries the *Q* allele on 5A yet is non-free-threshing due to tough adherent glumes (Muramatsu 1979; Simons et al. 2006) would suggest that domesticated emmer probably carries the *Tg2^{2B}* allele. No *Tg* gene has been

described on the A genome of polyploid wheat or from the A-genome progenitor *T. urartu*.

Kerber and Dyck (1969) first described the tenacious glume trait in wheat and attributed the character to an incompletely dominant gene TgI^{2D} . Early cytogenetic work placed TgI^{2D} on the short arm of chromosome 2D (Kerber and Rowland 1974), and more recent molecular mapping experiments have validated the position of TgI^{2D} on chromosome arm 2DS (Jantasuriyarat et al. 2004; Nalam et al. 2007; Sood et al. 2009; Table 18.1). Sood et al. (2009) demonstrated that TgI^{2D} and *sog* are not homoeologous suggesting that tgI^{2D} and *Sog* arose from independent mutations at non-orthologous loci. Therefore, it is possible that $Tg2^{2B}$ is homoeologous with either *Sog* on 2A^m from einkorn wheat or TgI^{2D} , but not both. Additional analysis by Nalam et al. (2007) suggested that two closely linked, possibly paralogous, TgI^{2D} loci exist on 2DS. Further work is necessary to validate this work and to clarify the genetic relationships among the *Tg/sog* loci.

18.6.3 The *Q* Loci

In addition to the *Tg* loci, the *Q* locus on wheat chromosome 5A also controls the free-threshing character. A mutation in the primitive q^{5A} allele led to the formation of the partially dominant Q^{5A} allele, which results in free-threshing seed (Fig. 18.3, Table 18.1). *Tg* is epistatic to *Q* because plants that have TgI^{2D} and Q^{5A} are not free-threshing (Kerber and Rowland 1974), but both the tgI^{2D} and Q^{5A} alleles are necessary to confer the free-threshing trait (Fig. 18.3). The *Q* locus has been an intriguing subject of study for the past 100 years due to the fact that it affects a repertoire of traits. Plants that have *br* and *tg* alleles but lack the Q^{5A} allele (possess the q^{5A} allele) are non-free threshing, have a semi-brittle rachis, a spike that is lax and primitive in appearance (speltoid), somewhat tenacious glumes that adhere to the seed, and are taller, flower earlier, and differ in yield compared to plants that harbor the Q^{5A} allele (Watkins 1940; Mackey 1954, 1966; Sears 1956; Muramatsu 1963, 1979, 1985, 1986; Singh 1969; Kato et al. 1999, 2003; Faris and Gill 2002; Faris et al. 2003, 2005; Simons et al. 2006; Zhang et al. 2011). Therefore, Q^{5A} pleiotropically affects numerous domestication-related and agronomically important traits.

Of the major domestication genes in wheat, *Q* is the only one that has been cloned so far (Faris et al. 2003; Simons et al. 2006). *Q* is a member of the AP2 family of transcription factors. Related members include *APETALA2*, which controls flower and seed development in *Arabidopsis* (Jofuku et al. 1994) and *indeterminate spikelet1 (ids1)*, which governs spikelet meristem fate in maize (Chuck et al. 1998). Homologs also exist in the other grasses such as rice, barley, *Brachypodium*, and sorghum (Simons et al. 2006; Faris et al. 2008), but functions have not been ascribed to the gene in these species. So far, wheat is the only plant species known for which this AP2-like gene has been recruited for domestication.

DNA sequence analysis of Q^{5A} and q^{5A} alleles from various wheat species of different ploidy levels validated the notion that q^{5A} is the more primitive allele and

that Q^{5A} formed once as the result of a mutation (Simons et al. 2006). This work also demonstrated that the A-genome diploids *T. monococcum* ssp. *monococcum* and *T. urartu* as well as the AB tetraploids *T. turgidum* spp. *dicoccoides* and *dicoccum*, and hexaploid *T. aestivum* spp. *macha* and *spelta* (European) all possess the primitive q^{5A} alleles. The free-threshing wheats including the tetraploids *T. turgidum* spp. *polonicum*, *carthilicum*, and *durum*, the free-threshing hexaploid *T. aestivum* ssp. *aestivum*, and also the non free-threshing *T. aestivum* ssp. *spelta* (Asian) shown by Luo et al. (2000) to have Q^{5A} , were all verified to have the Q^{5A} allele. The latter finding further supports the notion that Asian spelta has an origin different from that of European spelta and that it probably possesses either TgI^{2D} or $Tg2^{2B}$, which would mask the effects of the free-threshing Q^{5A} allele.

Comparative sequence analysis revealed only one conserved structural difference between Q^{5A} and q^{5A} alleles at the protein level: all q^{5A} alleles harbored a valine at amino acid position 329 whereas all Q^{5A} alleles had an isoleucine (Simons et al. 2006). The V₃₂₉ to I mutation was found to lead to an abundance of homodimer formation by the Q^{5A} protein. Transcription levels of Q^{5A} were also found to be more than twice the level of q^{5A} , and it is hypothesized that protein homodimerization may be a mechanism of self-regulation. Indeed, increased transcription levels of Q^{5A} and its effects on the domestication-related phenotypes were clearly demonstrated in transgenic plants (Simons et al. 2006), which confirmed the reports of Muramatsu (1963) regarding the dosage effects of Q using cytogenetic stocks. Another potential mechanism of Q regulation is a microRNA172 binding site in exon 10, which could mimic the regulation of *APETALA2* in *Arabidopsis* at the level of translation.

Clearly more work is needed to understand the mechanisms responsible for Q gene regulation, and its interactions with other genes. For example, Jantasuriyarat et al. (2004) evaluated a population of recombinant inbred lines derived from a cross between a hexaploid wheat variety and a synthetic hexaploid wheat line, which was created by crossing *T. turgidum* ssp. *durum* with *Ae. tauschii* followed by chromosome doubling thereby essentially repeating the hybridization event that led to formation of hexaploid wheat (a practice that is routinely done for the exploitation of desirable traits from the progenitor species). QTL analysis of the threshability trait revealed a locus with strong effects on the short arm of chromosome 2D in the vicinity of TgI^{2D} , as was expected because the synthetic parent is non free-threshing due to TgI^{2D} from the *Ae. tauschii* parent. However, a QTL on 5A near the Q locus was also detected, which was unexpected due to the assumption that both parents possessed the Q^{5A} allele. The effects of the Q^{5A} locus were thereby attributed to possible allelic variation within Q^{5A} . Similar work using a population of doubled haploid lines derived from different synthetic and cultivated wheat parents and subsequent analysis of QTL associated with threshability revealed essentially the same results (Faris JD, Chu CG, Friesen TL, Xu SS, unpublished). But, in this case we sequenced the Q^{5A} alleles from both parents and found no variation within the gene coding sequences. This might suggest that the effects of the Q locus on threshability in synthetic hexaploid-derived populations could be due to variation in Q^{5A} gene expression, possibly influenced by TgI^{2D} or other genes. Work is currently underway to gain further understanding of this phenomenon.

Homoeologous *q* loci on chromosomes 5B and 5D were long thought to exist and the work of Simons et al. (2006) proved it so. Zhang et al. (2011) conducted studies to better understand the evolution, organization, and function of homoeologous q^{5B} and q^{5D} alleles and their relationships with Q/q^{5A} alleles. Their analysis revealed that Q/q gene sequences were highly conserved among A, B, and D genomes in hexaploid wheat, the A and B genomes in tetraploid wheat, and the A, S, and D genomes in the diploid progenitors, but a duplication of the *q* gene prior to radiation of the diploid progenitors some 5.8 million BP was followed by the selective loss of one copy from the A genome progenitor and the other copy from the B, D, and S genomes. *Ae. tauschii* as well as hexaploid wheat possess intact and functional q^{5D} alleles, and functional and phenotypic analysis indicated that q^{5D} contributes to the suppression of the speltoid syndrome just as Q^{5A} , but to a lesser degree. *Ae. speltoides* possessed an intact and functional q^{5S} allele, but q^{5B} became a pseudogene upon formation of the AB tetraploid. However, q^{5B} still contributes to domestication-related traits through mechanisms of homoeoallele co-regulation in a rather complex manner. Therefore, while the mutation that gave rise to Q^{5A} from q^{5A} was a major factor in domesticating wheat, the contributions of q^{5B} and q^{5D} through polyploidization are also recognized as relevant contributors.

18.6.4 The Evolution of Free-threshing Wheats

A set of rather profound circumstances occurred in the formation of free-threshing polyploid wheat, especially given that the evolutionary steps involved mutations at three major loci (*Br*, *Tg*, and *q*) and two allopolyploidization events. Some of the events and transitions are well understood, while others are not, but clearly both genetics and the archaeological record indicate that all these events happened during an “evolutionary burst” that occurred relatively quickly over a period of less than a few thousand years. The first transition was of course the acquisition of a non-brittle rachis in *T. turgidum*. This required mutations in both $Br1^{3A}$ and $Br1^{3B}$ in wild emmer leading to the formation of domesticated emmer (Figs. 18.2, 18.3). Although domesticated emmer has a non-brittle rachis, it is not free-threshing because it carries the q^{5A} allele and most likely $Tg2^{2B}$ (Fig. 18.2). Free-threshing tetraploid wheat evolved from domesticated emmer by way of the mutation in q^{5A} that gave rise to Q^{5A} and also likely $Tg2^{2B}$ to $tg2^{2B}$ thus forming the free-threshing tetraploid, which is the genotype of today’s modern durum wheat. Not only did the mutation to Q^{5A} result in the free-threshing character, it also conferred a fully tough rachis.

It is most likely that a free-threshing tetraploid was involved in the amphiploidization event that gave rise to the first hexaploid (Fig. 18.2). However, the first hexaploid was non free-threshing due to acquisition of the $Tg1^{2D}$ gene from *Ae. tauschii*. It also likely acquired genes from *Ae. tauschii* for brittle rachis, possibly $Br1^{3D}$, $Br2^{3D}$, or both, and had the genotype $br1^{3A}br1^{3A}br1^{3B}br1^{3B}Br1^{3D}Br1^{3D}Br2^{3D}Br2^{3D}tg2^{2B}tg2^{2B}Tg1^{2D}Tg1^{2D}Q^{5A}Q^{5A}q^{5B}q^{5B}q^{5D}q^{5D}$. Therefore, mutations in $Tg1^{2D}$ and *Br* gene(s) obtained from *Ae. tauschii* were necessary in the transition of this

“Asian spelta-like” wheat to modern free-threshing *T. aestivum* ssp. *aestivum*. If the AB genome donor were domesticated emmer, the first hexaploid would have had to undergo mutations in q^{5A} and probably $Tg2^{2B}$ in addition to $Tg1^{2D}$ and possibly $Br1^{3D}$ and/or $Br2^{3D}$ to give rise to *T. aestivum* ssp. *aestivum*. Under this unlikely scenario, the free threshing tetraploids would have evolved later than free-threshing hexaploids through the acquisition of Q^{5A} from *T. aestivum* ssp. *aestivum*.

The archaeological record does not shed much light on the unknowns of these scenarios. Free-threshing tetraploids and hexaploids appear at the same time, shortly after the appearance of domesticated emmer, so the exact AB tetraploid progenitor and the answer to whether Q^{5A} first arose in the tetraploids or the hexaploids is yet unknown. Also, no primitive non free-threshing hexaploids precede the free-threshing hexaploids in the archaeological record. Genetic studies have discovered the origin of European spelta, and indicate that it is a derivative rather than a progenitor of *T. aestivum* ssp. *aestivum*, but no genetic evidence allows conclusions on whether or not Asian spelta is an ancient wheat or the result of a recent hybridization.

If the hybridization event that gave rise to hexaploid wheat occurred when *Ae. tauschii* came into contact with a field of free-threshing tetraploid wheat, the result would have been a hulled hexaploid growing in a field of free-threshing tetraploid wheat. As Nesbitt (2001) points out, early farmers would have processed hulled wheats different from free-threshing wheats, and as a result, hulled wheat grown in a free-threshing field would have been under strong selection pressure to become free-threshing itself. Thus, the transition would have occurred rapidly, which would explain large absence of spelt wheat from the Middle East. If domesticated emmer hybridized with *Ae. tauschii* to form the first hexaploid, mutations would have had to occur at q^{5A} , two *Tg* loci, and possibly two *Br* loci at the same time to explain the absence of spelt from the archaeological record. And, under this unlikely scenario, a free-threshing tetraploid would not have yet existed because we know that the mutation forming Q^{5A} occurred only once, indicating that Q^{5A} (and possibly $tg2^{2B}$) would have been acquired by an AB-tetraploid via gene flow from the hexaploid. This gene-flow event would have had to occur very rapidly to explain the appearance of free-threshing tetraploids and hexaploids at the same time. Whatever the scenario, the transitions necessary for free-threshing tetraploid and hexaploid wheat to evolve must have occurred very rapidly. Cloning and further genetic analysis of *Tg* and *Br* loci along with *Q/q* genes should provide more definitive answers to these unknowns in the future.

18.7 Wheat Evolution Under Cultivation

18.7.1 Capture of Genetic Variability

Domesticated emmer wheat arose through very few mutations in their wild progenitor subspecies, making the first domesticated form genetically very similar to the corresponding wild form. Even though domesticated emmer is by and large a

self-pollinated species, hybrids between wild and domesticated tetraploids are fully fertile and their chromosomes readily pair and recombine. Early farmers grew domesticated emmer in mixtures with wild emmer for a long period of time allowing ample opportunity for gene flow to occur leading to increased genetic variability, formation of sub-populations, and reduction of the founder effect (Luo et al. 2007). Gene flow has continued to occur even after domesticated emmer largely replaced wild emmer and other tetraploid wheats have come to exist such as today's economically important durum wheat. Many of the subspecies come into contact with each other at the edges of fields of cultivated durum or even hexaploid wheat and form hybrid swarms resulting in gene flow from wild to cultivated forms and vice versa (Dvorak et al. 2006; Syouf et al. 2006).

The founder effect for hexaploid wheat is much larger compared to emmer wheat because very few hybridization events between *T. turgidum* and *Ae. tauschii* occurred resulting in a larger magnitude of genetic drift. This and the fact that hexaploid was relatively isolated genetically because it hybridizes less frequently with its progenitors resulted in a much narrower genetic base and reduced variability compared to domesticated tetraploid wheats. However, the level of diversity is less restricted in the A and B genomes compared to the D genome, and in support of this, studies have shown that hybrid swarms involving wild emmer and common wheat exist and that gene flow from the former to the latter has occurred (Dvorak et al. 2006). In line with this, there is strong evidence that the hexaploid *T. aestivum* spp. *macha* and *spelta* (European), which have primitive traits, originated from interspecific crosses involving hexaploids and tetraploids (Bertsch 1943; MacKey 1966; Dvorak and Luo 2001; Blatter et al. 2002, 2004; Yan et al. 2003), and the tetraploid *T. turgidum* ssp. *carthlicum* probably arose through gene flow involving hexaploid wheat (Kuckuck 1979). Therefore, these new subspecies may have resulted from hybrid swarms as well.

18.7.2 Further Domestication Under Cultivation

Increased seed size was also an important domestication trait, and was selected for very early in the first years of wheat cultivation, probably before the acquisition of a non-brittle rachis (Fuller 2007). Seed size is under polygenic control and genes and QTLs governing this trait have been located on many wheat chromosomes (Peng et al. 2011). Following the transition to a non-brittle rachis, other traits acquired by newly domesticated emmer wheat included non-dormant seeds with uniform germination, yield, and probably more erect plants (Feldman 2001). As with seed size, many of these traits are under polygenic control and multiple QTLs associated with these traits have been reported (Peng et al. 2003, 2011). The spread of wheat cultivation throughout Europe, Africa, and Asia also required wide adaptation to the different environments, which included alterations in flowering time and growth habit.

According to Feldman (2001), a second phase of evolution under cultivation involved a lengthy period of time where continuous selection occurred for various

traits in fields consisting of mixtures of different genotypes and likely even different species of different ploidy levels. These mixtures of different landraces provided some protection against disease epidemics and probably some environmental hazards such as drought, extreme heat, and flooding as well. Therefore, the cultivation of these mixtures provided some level of crop security and ensured yield stability. These mixtures also provided ample opportunity for interspecific hybridizations and gene flow, thus increasing genetic variability. However, they also created a competitive environment where plants that had more tillers and were taller with horizontal leaves would shade the competitors and thus have a selective advantage.

The dawn of modern breeding practices began about a century ago. Since that time, individual genotypes have been the unit of selection rather than mixtures, which has led to wheat fields consisting of single genotypes with no opportunity for interspecific genetic exchange to increase variability. However, plant breeding practices in the mid-20th century allowed the development of the first uniform high-yielding cultivars through the introgression of elite agronomic characters. A profound example of this was the replacement of tall varieties with semi-dwarf and dwarf varieties through the introduction of the *Rht* genes. This occurred during the “green revolution” under the leadership of Norman E. Borlaug at the International Maize and Wheat Improvement Centre (CIMMYT) in Mexico. Breeding practices today allow extensive and intricate manipulations of the wheat plant to be made with certainty as well as rather precise monitoring of introgressions and genetic exchanges resulting from artificial hybridizations through the use of modern genomics tools and resources.

18.7.3 *Generation of New Genetic Diversity*

Although hexaploid wheat captured a good amount of genetic diversity for the A and B genomes, the genetic bottlenecks created by domestication resulted in the increased frequency of many adapted alleles but also the loss of other potentially useful ones. This limited amount of diversity reduces the potential for wheat to further adapt to changing environments. However, genomic analyses of wheat and its relatives over the past decade has revealed the wheat genomes in a polyploid state are dynamic and undergo the generation of new variation in the form of deletions, insertions, and point mutations created by repetitive elements affecting genes and regulatory elements (Dubcovsky and Dvorak 2007). Such alterations are well tolerated in wheat due to its polyploid buffering capacity. These alterations in gene coding regions create intergenic polymorphisms, which further increases genetic variability among the genomes and may lead to subtle changes in expression due to gene dosage differences. Alternatively, homoeologous sets of genes in a polyploid state are given the opportunity to become altered, for example, through sub-functionalization (becoming limited in function) or neo-functionalization (acquiring a new function). A good example of this is *Q* and its homoeoalleles because under polyploidization, Q^{5A} became hyper-functional, q^{5B} underwent pseudogenization, and q^{5D} became

sub-functionalized (Zhang et al. 2011). Therefore, to some extent, polyploid wheat is able to create a level of its own diversity due to the plasticity of its genomes and cooperation between genetic mutations created by repetitive elements and tolerance through polyploid buffering. Nevertheless, we need to make concentrated efforts to preserve a large reservoir of wild relatives and landraces through germplasm collections and gene banks so that we have the ability to introduce additional variability when needed.

18.8 Future Needs

Today's wheat varieties are far superior in yield, quality, biotic and abiotic resistance, and overall agronomic performance than even those that were considered elite just a few decades ago. However, more profound advances in these and other traits must be made in the near future if we are to feed the world's growing population. Significant advances in our understanding of the biology of the wheat plant must be achieved, and this can be initiated through the cracking of the genetic code of wheat and characterizing the structure and function of critical genes (see chapter by C. Feuillet). Obtaining the genome sequence of polyploid wheat and its relatives will also allow scientists to conduct studies that yield important information that may help answer unknowns regarding wheat evolution and domestication. This in turn will help wheat scientists to better cope with global climate change by allowing the recovery and exploitation of valuable alleles that exist in key wheat relatives. Under a continuously changing climate, the world may well see a need for another agricultural revolution and neo-domestication of wheat in order to meet future demands.

References

- Aaronsohn A (1910) Agricultural and botanical explorations in Palestine. Bull Plant Industry, US Dept Agriculture, Washington, DC No. 180:1–63
- Bertsch F (1943) Der Dinkel. Landw Jahrbuch 92:241–252
- Blake NK, Leffler BR, Lavin M, Talbert LE (1999) Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat. *Genome* 42:351–360
- Blatter RHE, Jacomet S, Schlumbaum A (2002) Spelt-specific alleles in HMW glutenin genes from modern and historical European spelt (*Triticum spelta* L.). *Theor Appl Genet* 104:329–337
- Blatter RHE, Jacomet S, Schlumbaum A (2004) About the origin of European spelt (*Triticum spelta* L.): allelic differentiation of the HMW glutenin B1-1 and A1-2 subunit genes. *Theor Appl Genet* 108:360–367
- Cao W, Scoles GJ, Hucl P (1997) The genetics of rachis fragility and glume tenacity in semi-wild wheat. *Euphytica* 94:119–124
- Chalupska D, Lee HY, Faris JD et al (2008) *Acc* homoeoloci and the evolution of the wheat genomes. *Proc Natl Acad Sci USA* 105:9691–9696
- Chen Q-F, Yen C, Yang J-L (1998) Chromosome location of the gene for brittle rachis in the Tibetan weed race of common wheat. *Genet Res Crop Evol* 45:21–25
- Chuck G, Meeley RB, Hake S (1998) The control of maize spikelet meristem fate by the *APETALA2*-like gene *indeterminant spikelet1*. *Genes Dev* 12:1145–1154

- de Moulins D (2000) Abu Hereyra 2: plant remains from the Neolithic. In: Moore AMT, Hillman GC, Legge AJ (eds) *Village on the Euphrates*. Oxford University Press, Oxford, pp 399–422
- Dixon A, Braun HJ, Kosina PP, Crouch J (2009) Wheat facts and futures. CIMMYT, Mexico
- Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866
- Dvorak J, Zhang HB (1990) Variation in repeated nucleotide sequences sheds light on the phylogeny of the wheat B and G genomes. *Proc Natl Acad Sci USA* 87:9640–9644
- Dvorak J, di Terlizzi P, Zhang H-B, Resta P (1993) The evolution of polyploid wheats: identification of the A genome donor species. *Genome* 36:21–31
- Dvorak J, Luo MC (2001) Evolution of free-threshing and hulled forms of *Triticum aestivum*: old problems and new tools. In: Caligari PDS, Brandham PE (eds) *Wheat taxonomy: the legacy of John Percival*. Linnean Society, London, pp 127–136 (Linnean Special Issue 3)
- Dvorak J, Luo MC, Yang ZL, Zhang HB (1998) The structure of the *Aegilops tauschii* gene pool and the evolution of hexaploid wheat. *Theor Appl Genet* 97:657–670
- Dvorak J, Akhunov ED, Akhunov AR et al (2006) Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat. *Mol Biol Evol* 23:1386–1396
- Fairbairn A, Asouti E, Near J, Martinoli D (2002) Macro-botanical evidence for plant use at Neolithic Catalhoyuk, south-central Anatolia, Turkey. *Veg Hist Archaeobot* 11:41–54
- Faris JD, Gill BS (2002) Genomic targeting and high-resolution mapping of the domestication gene *Q* in wheat. *Genome* 45:706–718
- Faris JD, Fellers JP, Brooks SA, Gill BS (2003) A bacterial artificial chromosome contig spanning the major domestication locus *Q* in wheat and identification of a candidate gene. *Genetics* 164:311–321
- Faris JD, Simons KJ, Zhang Z, Gill BS (2005) The wheat super domestication gene *Q*. *Wheat Info Serv* 100:129–148
- Faris JD, Zhang Z, Fellers JP, Gill BS (2008) Micro-colinearity between rice, *Brachypodium*, and *Triticum monococcum* at the wheat domestication locus *Q*. *Funct Integr Genomics* 8:149–164
- Feldman M (2001) Origin of cultivated wheat. In: Bonjean AP, Angus WJ (eds) *The world wheat book. A history of wheat breeding*. Lavoisier Publishing, Paris, pp 3–56
- Fuller DQ (2007) Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the Old World. *Ann Bot* 100:903–924
- Giles RG, Brown TA (2006) *GluDy* allele variations in *Aegilops tauschii* and *Triticum aestivum*: implications for the origins of hexaploid wheats. *Theor Appl Genet* 112:1563–1572
- Harlan JR, Wet MJ de, Price EG (1973) Comparative evolution of cereals. *Evolution* 27:3110–325
- Heun M, Schaefer-Pregl R, Klawan D et al (1997) Site of einkorn wheat domestication identified by DNA fingerprinting. *Science* 278:1312–1314
- Hillman GC (1978) On the origins of domestic rye—Secale cereal: the finds from Aceramic Can Hasan III in Turkey. *Anatolian Studies* 28:157–174
- Huang S, Sirikhachornkit A, Su X et al (2002) Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci USA* 99:8133–8138
- Jaaska V (1978) NADP-dependent aromatic alcohol dehydrogenase in polyploid wheats and their relatives. On the origin and phylogeny of polyploid wheats. *Theor Appl Genet* 53:209–217
- Jaaska V (1980) Electrophoretic survey of seedling esterases in wheats in relation to their phylogeny. *Theor Appl Genet* 56:273–284
- Jaaska V (1981) Aspartate aminotransferase and alcohol dehydrogenase isozymes: intraspecific differentiation in *Aegilops tauschii* and the origin of the D genome polyploids in the wheat group. *Plant Syst Evol* 137:259–273
- Jakubziner MM (1958) New wheat species. In: Jenkins BC (ed) *Proceedings of the first international wheat genetics symposium*. Winnipeg, pp 207–220

- Jantasuriyarat C, Vales MI, Watson CJW, Riera-Lizarazu O (2004) Identification and mapping of genetic loci affecting free-threshing habit and spike compactness in wheat (*Triticum aestivum* L.). *Theor Appl Genet* 108:261–273
- Jofuku KD, den Boer BGW, Van Montagu M, Okamuro JK (1994) Control of *Arabidopsis* flower and seed development by the homeotic gene *APETALA2*. *Plant Cell* 6:1211–1225
- Johnson BL (1968) Electrophoretic evidence on the origin of *Triticum zhukovskiyi*. In: Finlay KW, Shepherd KW (eds) *Proceedings of the Third International Wheat Genetics Symposium*. Canberra, Australia, pp 105–110
- Johnson BL, Dhaliwal HS (1976) Reproductive isolation of *Triticum boeoticum* and *Triticum urartu* and the origin of the tetraploid wheats. *Am J Bot* 63:1088–1094
- Kato K, Miura H, Sawada S (1999) QTL mapping of genes controlling ear emergence time and plant height on chromosome 5A of wheat. *Theor Appl Genet* 98:472–476
- Kato K, Sonokawa R, Miura H, Sawada S (2003) Dwarfing effect associated with the threshability gene *Q* on wheat chromosome 5A. *Plant Breed* 122:489–492
- Kerber ER (1964) Wheat: Reconstitution of the tetraploid component (AABB) of hexaploids. *Science* 143:253–255
- Kerber ER, Dyck PL (1969) Inheritance in hexaploid wheat of leaf rust resistance and other characters derived from *Aegilops squarrosa*. *Can J Genet Cytol* 11:639–647
- Kerber ER, Rowland GG (1974) Origin of the free threshing character in hexaploid wheat. *Can J Genet Cytol* 16:145–154
- Kihara H (1944) Discovery of the DD-analyzer, one of the ancestors of *Triticum vulgare*. *Agriculture and Horticulture* 19:13–14 (Tokyo)
- Kilian B, Ozkan H, Deusch O et al (2007) Independent wheat B and G genome origins in outcrossing *Aegilops* progenitor haplotypes. *Mol Biol Evol* 24:217–227
- Kimber G, Sears ER (1987) Evolution in the genus *Triticum* and the origin of cultivated wheat. In: Heyne EG (ed) *Wheat and wheat improvement*. 2nd edition. American Society of Agronomy, Madison, pp 154–164
- Kislev ME (1980) *Triticum parvicoccum*, the oldest naked wheat. *Isr J Bot* 28:95–107
- Kislev ME (1984) Emergence of wheat agriculture. *Paleorient* 10:61–70 (http://persee.fr/web/revues/home/prescript/article/paleo_0153-9345_1984_num_10_2_940)
- Kuckuck H (1959) Neuere Arbeiten zur Entstehung der hexaploiden Kulturweizen. *Z. Pflanzenzücht* 41:205–226
- Kuckuck H (1979) On the origin of *Triticum carthlicum* Neyski (= *Triticum persicum* Vav.). *Wheat Inf Serv* 50:1–5
- Lelley T, Stachel M, Grausgruber H, Vollmann J (2000) Analysis of relationships between *Ae. tauschii* and the D genome of wheat utilizing microsatellites. *Genome* 43:661–668
- Li WL, Gill BS (2006) Multiple pathways for seed shattering in the grasses. *Funct Integr Genomics* 6:300–309
- Liu Y-G, Tsunewaki K (1991) Restriction fragment length polymorphism (RFLP) analysis in wheat. II. Linkage maps of the RFLP sites in common wheat. *Jpn J Genet* 66:617–633
- Luo MC, Yang ZL, Dvorak J (2000) The *Q* locus of Iranian and European spelt wheat. *Theor Appl Genet* 100:602–606
- Luo MC, Yang ZL, You FM et al (2007) The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor Appl Genet* 114:947–959
- MacKey J (1954) Neutron and X-ray experiments in wheat and revision of the speltoid problem. *Hereditas* 40:65–180
- MacKey J (1966) Species relationship in *Triticum*. *Hereditas Supplement* 2:237–276
- Matsuoka Y, Nasuda S (2004) Durum wheat as a candidate for the unknown female progenitor of bread wheat: an empirical study with a highly fertile F₁ hybrid with *Aegilops tauschii* Coss. *Theor Appl Genet* 109:1710–1717
- McFadden ES, Sears ER (1946) The origin of *Triticum spelta* and its free-threshing hexaploid relatives. *J Hered* 37:81–89, 107–116

- Moore AMT, Hillman GC, Legge AJ (2000) The significance of Abu Hureyra. In: Moore AMT, Hillman GC, Legge AJ (eds) *Village on the Euphrates*. Oxford University Press, Oxford, pp 475–525
- Mori N, Ishi T, Ishido T et al (2003) Origins of domesticated emmer and common wheat inferred from chloroplast DNA fingerprinting. In: Pogna NE, Romano M, Pogna EA, Galterio G (eds) *Proceedings of the 10th International Wheat Genetics Symposium*, Paestum, Italy. Istituto Sperimentale per la Cerealicoltura, Rome, pp 25–28
- Muramatsu M (1963) Dosage effect of the *spelta* gene *q* of hexaploid wheat. *Genetics* 48:469–482
- Muramatsu M (1979) Presence of the *vulgare* gene, *Q*, in a dense-spike variety of *Triticum dicoccum* Schübl. Report of the Plant Germ-Plasm Institute, Kyoto University, No. 4: pp 39–41
- Muramatsu M (1985) Spike type in two cultivars of *Triticum dicoccum* with the *spelta* gene *q* compared with the *Q*-bearing variety *liguliforme*. *Jpn J Breed* 35:255–267
- Muramatsu M (1986) The *vulgare* super gene, *Q*: its universality in durum wheat and its phenotypic effects in tetraploid and hexaploid wheats. *Can J Genet Cytol* 28:30–41
- Nakai Y (1979) Isozyme variation in *Aegilops* and *Triticum*, IV. The origin of the common wheats revealed from the study of esterase isozymes in synthesized hexaploid wheats. *Jpn J Genet* 54:175–189
- Nalam VJ, Vales MI, Watson CJW et al (2006) Map-based analysis of genes affecting the brittle rachis character in tetraploid wheat (*Triticum turgidum* L.). *Theor Appl Genet* 112:373–381
- Nalam VJ, Vales MI, Watson CJW, Johnson EB et al (2007) Map-based analysis of genetic loci on chromosome 2D that affect glume tenacity and threshability components of free-threshing habit in common wheat (*Triticum aestivum* L.). *Theor Appl Genet* 116:135–145
- Nesbitt M (2001) Wheat evolution: integrating archaeological and biological evidence. In: Caligari PDS, Brandham PE (eds) *Wheat taxonomy: the legacy of John Percival*. Linnean Society, London, pp 37–59 (Linnean Special Issue 3)
- Nesbitt M, Samuel D (1996) From staple crop to extinction? The archaeology and history of hulled wheats. In: Padulosi S, Hammer K, Heller J (eds) *Hulled wheats, promoting the conservation and use of underutilized and neglected crops 4: proceedings of the first international workshop on hulled wheats*. Castelveccchio Pascoli, Tuscany, pp 41–100
- Nishikawa K (1974) Alpha-amylase isozymes and phylogeny of hexaploid wheat. In: Sears ER, Sears EMS (eds) *Fourth international wheat genetics symposium*, vol 1. University of Missouri, Columbia, pp 851–855
- Nishikawa K, Furuta Y, Wada T (1980) Genetic studies on alpha-amylase isozymes in wheat. III. Intraspecific variation in *Aegilops squarrosa* and birthplace of hexaploid wheat. *Jpn J Genet* 55:325–336
- Ozkan H, Brandolini A, Schafer-Pregl R, Salamini F (2002) AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in Southeast Turkey. *Mol Biol Evol* 19:1797–1801
- Ozkan H, Brandolini A, Pozzi C et al (2005) A reconsideration of the domestication geography of tetraploid wheats. *Theor Appl Genet* 110:1052–1060
- Peleg Z, Fahima T, Korol AB et al (2011) Genetic analysis of wheat domestication and evolution under domestication. *J Exp Bot* 62:5051–5061
- Peng JH, Ronin Y, Fahima T et al (2003) Domestication quantitative trait loci in *Triticum dicoccoides*, the progenitor of wheat. *Proc Natl Acad Sci USA* 100:2489–2494
- Peng JH, Sun D, Nevo E (2011) Domestication evolution, genetics and genomics in wheat. *Mol Breeding* 28:281–301
- Renfrew JM (1973) *Palaeoethnobotany—the prehistoric food plants of the Near East and Europe*. Methuen and Co. Ltd, London, pp 1–248
- Riley R, Unrau J, Chapman V (1958) Evidence on the origin of the B genome of wheat. *J Hered* 49:91–98
- Rodriguez J, Maestra B, Perera E, Diez M et al (2000) Pairing affinities of the B- and G- genome chromosomes of polyploid wheats with those of *Aegilops speltoides*. *Genome* 43:814–819

- Salse J, Chague V, Bolot S et al (2008) New insights into the origin of the B genome of hexaploid wheat: Evolutionary relationships at the *SPA* genomic region with the S genome of the diploid relative *Aegilops speltoides*. *BMC Genomics* 9:555
- Sarkar P, Stebbins GL (1956) Morphological evidence concerning the origin of the B genome in wheat. *Am J Bot* 43:297–304
- Sax K (1922) Sterility in wheat hybrids. II. Chromosome behavior in partially sterile hybrids. *Genetics* 7:513–552
- Sears ER (1956) The systematics, cytology and genetics of wheat. *Handb Pflanzenzücht*, 2nd Edition, 2:164–187
- Sharma HC, Waines JG (1980) Inheritance of tough rachis in crosses of *Triticum monococcum* and *T. aegilopoides*. *J Hered* 71:214–216
- Simonetti MC, Bellomo MP, Laghetti G et al (1999) Quantitative trait loci influencing free-threshing habit in tetraploid wheats. *Genet Res Crop Evol* 46:267–271
- Simons KJ, Fellers JP, Trick HN et al (2006) Molecular characterization of the major wheat domestication gene *Q*. *Genetics* 172:547–555
- Singh MP (1969) Some radiation induced changes at '*Q*' locus in bread wheat (*Triticum aestivum* L.). *Caryologia* 22:119–126
- Sood S, Kuraparthy V, Bai GH, Gill BS (2009) The major threshability genes soft glume (*sog*) and tenacious glume (*Tg*), of diploid and polyploid wheat, trace their origin to independent mutations at non-orthologous loci. *Theor Appl Genet* 119:341–351
- Syouf M, Abu-Irmaileh BE, Valkoun J, Bdour S (2006) Introgression from durum wheat landraces in wild emmer wheat (*Triticum dicoccoides* (Körn. ex Asch. et Graibner) Schweinf). *Genet Res Crop Evol* 53:1165–1172
- Taenzler B, Esposti RF, Vaccino P et al (2002) Molecular linkage map of einkorn wheat: mapping of storage-protein and soft-glume genes and bread-making quality QTLs. *Genet Res Camb* 80:131–143
- Tanno K, Willcox G (2006) How fast was wild wheat domesticated? *Science* 311:1886
- Tsunewaki K (1966) Comparative gene analysis of common wheat and its ancestral species. II. Waxiness, growth habit and awnedness. *Jpn J Bot* 19:175–229
- Watanabe N, Ikebata N (2000) The effects of homoeologous group 3 chromosomes on grain colour dependent seed dormancy and brittle rachis in tetraploid wheat. *Euphytica* 115:215–220
- Watkins AE (1940) The inheritance of glume shape in *Triticum*. *J Genet* 39:249–264
- Yan Y, Hsam SLK, Yu JZ et al (2003) HMW and LMW glutenin alleles among putative tetraploid and hexaploid European spelt wheat (*Triticum spelta* L.) progenitors. *Theor Appl Genet* 107:1321–1330
- Zeder M (2008) Domestication and early agriculture in the Mediterranean Basin: Origin, diffusion, and impact. *Proc Natl Acad Sci USA* 105:11597–11604
- Zhang P, Friebe B, Gill BS (2002) Variation in the distribution of a genome-specific DNA sequence on chromosomes reveals evolutionary relationships in the *Triticum* and *Aegilops* complex. *Plant Syst Evol* 235:169–179
- Zhang ZC, Belcram H, Gornicki P et al (2011) Duplication and partitioning in evolution and function of homoeologous *Q* loci governing domestication characters in polyploid wheat. *Proc Natl Acad Sci USA* 108:18737–18742

Chapter 19

Molecular Evidence for Soybean Domestication

Kyujung Van, Moon Young Kim, Jin Hee Shin, Kyung Do Kim,
Yeong-Ho Lee and Suk-Ha Lee

Contents

19.1 Introduction	466
19.2 Archeological Evidence of Soybean Domestication	468
19.3 Domestication-related Phenotypic Traits in Soybean	469
19.4 Molecular Diversity Related to Soybean Domestication	471
19.4.1 Molecular Markers	471
19.4.2 Sequence Diversity	473
19.5 Model of Soybean Domestication History	475
19.6 Conclusions	477
References	478

Abstract Crop domestication is a good example of plant–human co-evolution. Seed gathering and human cultivation of crops were observed since the Neolithic period, as shown by archeological evidence. Numerous studies have been conducted to identify genes related to domestication. With the development of molecular techniques (molecular markers and next-generation sequencing) and bioinformatics, a greater understanding of crop domestication and improvement has been established, including the origins of crops, the numbers of independent domestication events, the molecular diversity of domestication-related traits (DRTs), and the selection pressures. A comparison of the genome sequences between wild species and cultivated crops may provide key information regarding the genetic elements involved in speciation and domestication. Therefore, sequencing projects of currently important crops and their wild relatives are in progress. Accordingly, whole genome sequencing of soybean could provide new knowledge about domestication of this important crop. In this review, we introduce the archaeological evidence of soybean domestication and summarize the DRTs in soybean populations of crosses between cultivated

S.-H. Lee (✉) · M. Y. Kim

Plant Genomics and Breeding Institute, Department of Plant Science and Research
Institute for Agriculture and Life Sciences, Seoul National University,
San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea
e-mail: sukhalee@snu.ac.kr

K. Van · J. H. Shin · K. D. Kim · Y.-H. Lee

Department of Plant Science and Research Institute for Agriculture and Life Sciences,
Seoul National University, San 56-1, Sillim-dong, Gwanak-gu, Seoul 151-921, Korea

(*Glycine max*) and wild soybean (*G. soja*). Soybean domestication is discussed at the sequence level. The current hypothesis of soybean domestication considers that *G. max* was domesticated from *G. soja*. However, our previous work suggested that soybean was domesticated from the *G. soja/G. max* complex that diverged from a common ancestor of these two species of *Glycine*. This review explores soybean domestication history by focusing on nucleotide diversity using resequencing. Analysis of genes around DRTs at the population level may clarify the domestication history of soybean.

Keywords Cultivated soybean · Domestication · Domestication-related traits (DRTs) · Next-generation sequencing (NGS) · Wild soybean

19.1 Introduction

Domestication of plant and animal species is considered an evolutionary process that involves human intervention to select specific morphological and physiological traits (Purugganan and Fuller 2009). As human civilization has evolved, the rapid domestication of selected species was also accelerated. This is one of the important technological innovations in human history (Diamond 2002). Charles Darwin was also very interested in domestication when his theory on the origin of species through natural selection was formulated (Purugganan and Fuller 2009).

The interaction of humans with plants was the driving force behind crop domestication. Through archaeological and ethnographic studies of traditional farming societies and hunter-gatherers, the selections of early farmers (the first plant breeders) contributed to the domestication of some of the cultivated crop species (Doebley et al. 2006; Vaughan et al. 2007; Purugganan and Fuller 2009; Tang et al. 2010). These early farmers planted, harvested and reselected crops after they identified useful genetic traits from wild species. Centuries later, plant hybridization and Mendel's discovery of genetic laws provided much greater efficiency in crop breeding programs and helped to develop new crop varieties with desirable traits (Vaughan et al. 2007). Hajjar and Hodgkin (2007) suggested that a detailed genetic analysis between various domesticated crop species and their wild relatives could be useful for understanding the evolutionary relationships between crops and their wild progenitors. Since domestication resulted in reduced genetic diversity and the loss of some useful traits from wild progenitors, the wild relatives of crops should be conserved as valuable germplasm for future use in agriculture (Doebley 1989; Vaughan et al. 2007; Guo et al. 2010).

The main domesticated crops, such as rice, wheat, maize and tomato, were intensively studied by molecular geneticists. Two important reviews have summarized the genes and their phenotypes/traits that are related to crop domestication and improvement (Doebley et al. 2006; Gross and Olsen 2010). These data have provided useful resources for understanding the geographical origins of crops, the domestication process and artificial selection during the domestication, diversification and improvement of crops.

Although the study of the domestication of rice (*Oryza sativa*) is an active area of research, the archaeological record of its domestication and early history are still unclear (Doebley et al. 2006; Kovach et al. 2007). Some evidence suggested that *O. sativa* was domesticated from *O. rufipogon*, an Asian common wild rice. However, the two varietal groups in *O. sativa* (*indica* and *japonica*) were generated from different gene pools within *O. rufipogon* based on genome-wide studies of genetic variation. This multiple domestication of *O. sativa* became more complicated when domestication-related genes were characterized (Kovach et al. 2007). The recently proposed ‘combination model’ for rice domestication considers that the early *O. indica* and *O. japonica* cultivars were domesticated from divergent *O. rufipogon*, and the key domestication alleles were introgressed and resulted in the current domestication alleles in modern varieties (Sang and Ge 2007). This complex rice domestication theory can be solved by studies of molecular markers with various rice accessions including landraces and modern cultivars.

Soybean (*Glycine max*) is a valuable and economically important food crop due to the high levels of protein and oil in its seed (Kim et al. 2012). The genus *Glycine* has an Old World distribution and includes the annual self-pollinated subgenus *Soja*, the wild perennial subgenus *Glycine* Willd., and the subgenus *Bracteata* (Hymowitz 1970). Both wild soybean (*G. soja* Sieb. and Zucc.) and cultivated soybean (*G. max* (L.) Merr.) are in the subgenus *Soja*. *G. tomentella* Hayata and *G. falcata* Benth. are the main species of the subgenus *Glycine* Willd. The wild soybean is distributed in East Asia, including China, Taiwan, Russian Far East, the Korean Peninsula, and Japan (Boerma and Specht 2004). There is controversy over the origin or domestication site of soybean and several regions were suggested, such as southern China, the Middle Yellow River valley of central China, north-eastern China, or multiple sites simultaneously (Hymowitz 1970; Boerma and Specht 2004).

Wild soybean (*G. soja*) is the closest relative of modern soybean (Kim et al. 2010). Although *G. soja* and *G. max* are morphologically quite different, both have 20 chromosomes ($2n = 40$) and show genome duplication, which has an evolutionary impact on the structure of the soybean genome (Van et al. 2008). Since easy hybridization and normal meiotic chromosome pairing were observed between wild and cultivated soybeans, wild soybean is a valuable resource for novel genes and alleles for soybean cultivar development (Stupar 2010). However, domesticated soybeans show a severe ‘genetic bottleneck’ with a reduction of genetic diversity because domestication can alter genetic diversity (Guo et al. 2010; Tang et al. 2010). During soybean domestication, 50 % of the genetic diversity and 81 % of the rare alleles have been lost and 60 % of the genes showed significant changes in allele frequency (Hyten et al. 2006). Mapping traits related to soybean domestication have been studied with various kinds of germplasm including domesticated and wild relatives (Liu et al. 2007), but only a soybean gene for determinate growth habit has been characterized so far (Liu et al. 2010; Tian et al. 2010). More molecular studies of population genetics are needed for understanding soybean domestication.

In this chapter, we discuss soybean domestication with recent archeological evidence and domestication-related phenotypic traits. We describe how molecular tools

such as molecular markers and sequence variations can be used to understand soybean domestication. Whole genome studies using next-generation sequencing (NGS) technology are already opening new chapters regarding domestication.

19.2 Archeological Evidence of Soybean Domestication

Using plant remains from sites of human activity, archaeobotanical evidence provides data for understanding the initial evolution of domesticated plants (Fuller 2007). The development of systematic sampling for archaeobotanical remains has enabled large numbers of complete samples to be collected in many sites (Fuller 2002). Based on archaeological data, a classic model proposed four general stages of domestication: the hunting and gathering of wild plant food, wild plant food production (the beginning of cultivation), systematic cultivation and agriculture of domesticated plants (Harris 1998). These efforts yielded an increase in plant food production that sometimes became a surplus.

Limited data suggest that soybean (*G. max*) was present in East Asia sites as early as the third millennium BC. The seeds of these soybean were larger compared to those found in the later Bronze Age or Iron Age (Crawford and Lee 2003; Crawford 2005; Fuller 2007). Lee et al. (2007) evaluated 26 soybean (*Glycine* Sp.) specimens from the Early Longshan (3000–2500 BC) to the Erligang (1600–1400 BC) periods. The sizes of these soybean specimens were much smaller than the domesticated specimens from Early Bronze Age (Mumun) sites of 1400–1000 calibrated calendrical years (cal) BC in Korea (Crawford and Lee 2003; Lee et al. 2007). The soybean sizes of the Erlitou period (ca. 1900–1500 BC) at the Zaojiaoshu site were intermediate between *G. soja* and *G. max*. Soybean was reported to grow in nearly 30 sites in China from 7000 BC to 200 AD, whereas wild soybean now grows throughout north-central China (Hymowitz 1970). However, many studies suggested that seed size is not the only trait for discriminating between wild and domesticated soybean because soybean seed size may not be the only selection criterion used by ancient farmers (Crawford et al. 2005; Lee et al. 2007; Lee 2011).

Recent data suggested that Korea is one of several regions where local domestication of soybean occurred (Lee and Park 2006; Lee 2011). In central and south-eastern Korea during the Chulmum period (the same time span as Neolithic, 7500–3400 before present (BP) or 5500–1400 BC), archeological evidence supported the transition to food production, environmental impulse, and migration. Soybean seeds were found in Pyeonggeodong sites (35° 13' N, 128° 07' E) and their sizes were about half of the domesticated soybean size from the Mumun period (Crawford and Lee 2003; Lee 2011). The soybean seed sizes of Pyeonggeodong sites were almost the same size as those of the Late Longshan (ca. 4500–4000 BP) soybeans in China (Lee et al. 2007). A domesticated soybean variety was also discovered from the Middle Jomon sites in Kyushu (ca. 5300–4400 cal BP) (Lee 2011). Soybean found at Pyeonggeodong sites was identified as a wild species based on morphology, but these sites appeared to have great economic importance because of the dense concentration of soybean found.

Analysis of carbonized soybean seeds including *G. soja* and *G. max* suggested that soybean was widely cultivated in the Korean Peninsula since the Bronze Age (1500–300 BC). The identification of various types of soybean indicated a high genetic diversity of soybean germplasm in the Korean Peninsula (Lee and Park 2006). Considering that the Korean Peninsula was formed about 11,000 years ago in the Mid-Glacial Epoch, the wild-type soybean varieties were already present. It is suggested that they were independently selected and domesticated by the early residents. Thus, this archeological evidence strongly suggests that the Korean Peninsula is one of the origins of soybean domestication and cultivation (Lee and Park 2006).

19.3 Domestication-related Phenotypic Traits in Soybean

Most seed crops can be distinguished from their progenitors by domestication-related phenotypic traits, such as larger seeds, more robust plants, more determinate growth or increased apical dominance, and a loss of natural seed dispersal. These domestication-related traits (DRTs) enabled humans to easily harvest seeds (Doebley et al. 2006). The traits selected by early farmers contributed to shape the phenotype of current soybean cultivars. Studying the DRTs is useful to identify important genes in wild species, such as those related to protein content and disease resistance, which might have been lost by human selection during the domestication of soybean. Genetic studies have examined phenotypic differences between domesticated soybean and its wild progenitor (Table 19.1). Wild soybean was simply considered to be one of the breeding parents with useful traits, such as disease and pest resistance. Using the breeding populations derived from crosses between wild and cultivated soybeans, quantitative trait loci (QTLs) responsible for DRTs have been reported (Table 19.1).

In early studies on DRTs, significant associations were identified between restriction fragment length polymorphism (RFLP) markers and DRTs, such as leaf width, leaf length, stem diameter, canopy height, stem length, first flower (R1), seed pod maturity (R8) and seed-fill (R1 to R8) (Keim et al. 1990a). Five independent RFLP markers associated with hard seededness were identified in a segregating population from a *G. max* and *G. soja* cross (Keim et al. 1990b). Because resistance to pod dehiscence is a favorable trait prior to harvest, intensive work has identified 17 QTLs for pod dehiscence. Nine out of 17 QTLs are positioned on soybean chromosome 16 (Table 19.1). Liu et al. (2007) surveyed soybean QTLs for many other DRTs, such as early flowering, twinning habit and seed weight, and concluded that most soybean DRTs were controlled by one or two major QTLs. This suggested that useful genes from wild soybean can be introgressed into cultivated soybean because major QTLs for DRTs were randomly distributed on soybean chromosomes.

Fine mapping of genes related to DRTs would be helpful for cloning these genes and understanding the molecular basis of domestication-related changes (Vaughan et al. 2007). Although many studies have mapped the traits associated with soybean domestication, only one trait for determinate growth habit has been characterized in detail at the genome level (Liu et al. 2010; Tian et al. 2010). Genome analysis by NGS technology and comparisons among different *Glycine* species are likely to help in identifying genes related to soybean domestication.

Table 19.1 List of soybean domestication-related traits^a in populations derived from *Glycine max* (cultivated soybean) and *Glycine soja* (wild soybean)

Trait	Gene	Chromosome	Marker	Position (cM)	R ² value	References	
Determinate habit	<i>GmTfl1</i>	19	Sat_099	78.2	72.5	Tian et al. (2010)	
						Liu et al. (2010)	
Twinning habit		1	Satt408	106.7	9.2	Liu et al. (2007)	
		2	Satt546	87.2	10.1	Liu et al. (2007)	
		18	Satt235	21.9	26.4	Liu et al. (2007)	
Number of nodes		5	Satt042	82.5	9.9	Liu et al. (2007)	
		18	Satt235	21.9	14.0	Liu et al. (2007)	
Pod dehiscence		2	A725	8.6	6.6	Bailey et al. (1997)	
		2	Satt350	15.9	5.1	Kang et al. (2009)	
		15	Sat_124	15.9	9.6	Liu et al. (2007)	
		15	cr274_1	45.0	7.3	Bailey et al. (1997)	
		15	BLT049_5	46.3	6.0	Bailey et al. (1997)	
		15	B124_3	47.5	4.5	Bailey et al. (1997)	
		15	cr324_1	54.2	5.8	Bailey et al. (1997)	
		16	B074_1	17.4	19.0	Bailey et al. (1997)	
		16	B166_1	27.6	23.0	Bailey et al. (1997)	
		16	Satt215	44.1	46.0	Kang et al. (2009)	
		16	Sat_093–Sat_366	46.1–52.8		Funatsuki et al. (2006)	
		16	ATG/CCG270	56.9 ^b	21.8	Liu et al. (2007)	
		16	B122_1	57.2	44.4	Bailey et al. (1997)	
		16	K375_1	68.3	3.5	Bailey et al. (1997)	
	Hard seededness		16	cr392_1	80.2	6.0	Bailey et al. (1997)
			16	ATG/CCG270	44.1	21.8	Liu et al. (2007)
		19	A489_1	95.4	5.7	Bailey et al. (1997)	
		8	I	48.8	32.0	Keim et al. (1990a)	
		8	T153_1	43.9	34.0	Keim et al. (1990a)	
		8	A111_1	67.3	11.0	Keim et al. (1990a)	
		6	ACG/CGC380	107.3 ^b	16.5	Liu et al. (2007)	
		2	Satt459	118.6	46.3	Liu et al. (2007)	
		2	K411_1	119.3	13.0	Keim et al. (1990a)	
		19	G173_1	86.6	15.0	Keim et al. (1990a)	
		3	K418_1	30.4	12.0	Keim et al. (1990a)	
		3	R022_1	16.7	11.0	Keim et al. (1990a)	
Maximum internode length			18	Satt235	21.9	17.0	Liu et al. (2007)
			19	ATC/CCG315	71.0 ^b	16.0	Liu et al. (2007)
Seed yield			5	Satt511	94.2	12.0	Li et al. (2008)
			14	Satt066	78.8		Concibido et al. (2003)
100 seed weight		14	Satt304	65.6	10.0	Li et al. (2008)	
		17	Satt154	57.1	19.3	Liu et al. (2007)	
		12	Satt541	53.3	14.0	Li et al. (2008)	
		7	AGA/CAG540/560	58.6 ^b	10.8	Liu et al. (2007)	
		10	Sat_274	107.6	11.3	Liu et al. (2007)	

Table 19.1 (continued)

Trait	Gene	Chromosome	Marker	Position (cM)	R ² value	References
First flower (R1)	<i>E1</i>	6	K365_1	120.4	21.0	Keim et al. (1990b)
	<i>E1</i>	6	K474_1	123.8	23.0	Keim et al. (1990b)
	<i>E1</i>	6	K474_2	124.0	21.0	Keim et al. (1990b)
Seed pod maturity (R8)		4	K472_1	53.9	18.0	Keim et al. (1990b)
		6	K365_1	120.4	21.0	Keim et al. (1990b)
		6	K474_1	123.8	18.0	Keim et al. (1990b)
		6	K474_2	124.0	21.0	Keim et al. (1990b)
		1	R013_2	38.1	20.0	Keim et al. (1990b)
		13	Satt335	77.7	10.0	Li et al. (2008)
Seed filling period (R1 to R8)		3	Satt584	38.0	9.0	Li et al. (2008)
		8	A111_1	67.3	18.0	Keim et al. (1990b)
		13	Satt335	77.7	12.0	Li et al. (2008)
		20	Satt330	77.8	10.0	Li et al. (2008)
Lodging		19	Satt284	38.2	15.0	Li et al. (2008)
		3	Satt549	70.6	10.0	Li et al. (2008)
		8	A111_1	67.3	24.0	Keim et al. (1990b)
Leaf width		13	K390_1	40.4	17.0	Keim et al. (1990b)
		2	K411_1	119.3	16.0	Keim et al. (1990b)
Leaf length		1	R013_2	38.1	18.0	Keim et al. (1990b)
		1	K478_1	34.9	19.0	Keim et al. (1990b)
Stem diameter		19	G173_1	86.6	24.0	Keim et al. (1990b)
		19	K385_1	101.3	17.0	Keim et al. (1990b)
		19	R201_1	102.0	17.0	Keim et al. (1990b)
Canopy height		13	K390_1	40.4	16.0	Keim et al. (1990b)
		1	R013_2	38.1	20.0	Keim et al. (1990b)

^a All traits, their positions, and values were acquired from Soybase (<http://soybase.org>)

^b Marker positions from the corresponding references

19.4 Molecular Diversity Related to Soybean Domestication

Molecular diversity analysis enabled the mapping of domestication-related genes and provided information on the effects of selection and domestication. As molecular and sequencing technologies rapidly developed, many different accessions could be easily analyzed for nucleotide polymorphisms and molecular genetic markers.

19.4.1 Molecular Markers

Domestication causes a decline in the genetic diversity of crop plants because only selected fruits and seeds will be propagated in the next generation (Doebley et al.

1989). As a result, domestication traits and genes will vary depending on the crops (Purugganan and Fuller 2009). Several domestication genes in plants have been identified through molecular genetic studies, and those genes were associated with human cultural preferences, such as higher yield and growth habits (Gunter 2008; Jin et al. 2008; Vielle-Calzada et al. 2009). Domestication creates genetic bottlenecks that can affect genetic diversity and allele frequencies and can eliminate rare alleles in domesticated populations (Doebley et al. 2006; Hyten et al. 2006).

The genetic diversity before and after domestication can be measured by comparing a crop and its wild relative, which may represent the pre-domestication population (Doebley et al. 2006). Crop wild relatives (CWRs) are wild plants closely related to domesticated crops and may include the crop progenitors (Hajjar and Hodgkin 2007). They provide genetic variation that may not exist among cultivated populations. Population genetics using CWRs and crops allows the identification of alleles related to historical events (Nielsen 2005). For example, in rice (*Oryza sativa*), domestication-related genes have been detected by genetic studies using the CWRs, such as *O. rufipogon*, *O. nivara*, *O. glumaepetula* and *O. meridionalis* (Sobrizal et al. 1999; Xiong et al. 1999; Cai and Morishima 2000; Nagai et al. 2002; Li et al. 2006). In maize (*Zea mays*), the analysis of nucleotide diversity between domesticated maize and its wild ancestor teosinte (*Z. mays* subsp. *mexicana* and subsp. *parviglumis*) has identified a selective sweep and loci with no genetic diversity in cultivated maize populations (Nielsen 2005; Vielle-Calzada et al. 2009).

Molecular markers have been used for exploring genetic diversity because they are highly polymorphic and not affected by the environment (Andersen and Lubberstedt 2003; Schulman 2007). Soybean genetic diversity in *G. max* and *G. soja* have been studied with various markers, including RFLPs, randomly amplified polymorphic DNAs (RAPDs), simple sequence repeats (SSRs), amplified fragment length polymorphisms (AFLPs) and single nucleotide polymorphisms (SNPs). These studies revealed considerably higher genetic diversity in *G. soja* than in *G. max*, and they identify a domestication bottleneck (Xu et al. 2002; Guo et al. 2010; Li et al. 2010; Jun et al. 2011). Li et al. (2010) assessed genetic diversity with SSR and SNP markers, and reported significantly higher allelic richness, gene diversity and allele numbers in *G. soja* than in *G. max*. For example, a total of 1807 alleles were observed from 92 *G. soja*, whereas a smaller number of 1473 alleles were identified from much larger sample size of 279 *G. max* (Li et al. 2010). Furthermore, landrace soybeans exhibited only 41.9% of the allelic diversity found in wild soybean, although the percentage of heterozygosity in landraces was 14.7% lower than wild soybean (Guo et al. 2010). Based on the coalescent simulation, the severity of the domestication bottleneck in soybean (2.0) is quite moderate compared to rice (0.2 for *japonica* and 0.5 for *indica*) but higher than that of maize (4.0–5.0) (Buckler et al. 2001; Zhu et al. 2003; Tenaillon et al. 2004; Hyten et al. 2006; Guo et al. 2010; Li et al. 2010). Jun et al. (2011) found that the genetic diversity in *G. soja* was higher than in *G. max* using SSR marker analysis with 192 *G. soja* and 104 *G. max* accessions. Strong genetic differentiation between *G. soja* and *G. max* was observed in SSRs near QTLs associated with seed protein content, suggesting that selective breeding for this trait of great agronomic importance played an important role in soybean domestication events.

There were clear differences between *G. soja* and *G. max* at the genetic diversity level, representing two distinct germplasm pools (Guo et al. 2010). Within a taxon, it was shown that the clustering of accessions was related to the geographical location (Xu et al. 2002; Lee et al. 2008; Li et al. 2008; Li et al. 2010). However, the origin of domestication is still unclear. A single domestication origin of cultivated soybean has been hypothesized based on molecular marker and phylogenetic analyses (Dong et al. 2004; Guo et al. 2010; Li et al. 2010), whereas some analyses supported multiple origins from different wild gene pools (Xu et al. 2002; Xu and Gai 2003). The different conclusions may have been caused by different markers and samples analyzed (Guo et al. 2010). Although molecular markers are good tools for genetic analysis such as QTL mapping and marker-assisted selection (MAS), the use of the whole genome sequence is expected to increase the resolution of studies for genetic diversity and domestication (Jackson et al. 2011).

Polymorphisms of molecular markers allowed QTL mapping for many traits including DRT (Ross-Ibarra 2005). QTL mapping for DRTs using a population derived from a cross between a crop and its wild relative has been conducted in crops such as aubergine (Doganlar et al. 2002), maize (Doebley et al. 1990; Doebley and Stec 1993; Lauter and Doebley 2002), pearl millet (Poncet et al. 2002), rice (Xiong et al. 1999; Cai and Morishima 2002), sunflower (Burke et al. 2002), tomato (Grandillo and Tanksley 1996; Frary et al. 2000), common bean (Koinange et al. 1996) and wheat (Peng et al. 2003). QTL mapping followed by map-based cloning in the recombinant mapping population allowed the identification of domestication genes (Gross and Olsen 2010). Many DRTs were identified in soybean using RFLP markers (Keim et al. 1990a, b). Molecular mapping of recombinant inbred lines (RILs) resulting from *G. max* and *G. soja* was performed for nine DRTs (Liu et al. 2007). Most of the traits such as flowering, determinate habit, twining habit, pod dehiscence, seed weight and hard seededness were found to be controlled by one or two major QTLs (Liu et al. 2007). Although QTL mapping results for DRTs could differ among populations, this information would be a good resource for the identification of domestication-related genes. Recently, a gene correlating with the determinate habit was identified among the nine DRTs and is described below.

19.4.2 Sequence Diversity

The sequence diversity of soybeans has been studied using SNPs and analysis of insertion/deletion (indels) among cultivated and wild soybeans. Intensive SNP discovery using coding or non-coding soybean sequences was first reported for North American soybean populations (Zhu et al. 2003). In 25 diverse soybean genotypes, nucleotide diversity (θ) was observed to be 0.00053 and 0.00111 in coding and non-coding perigenic DNA, respectively. These estimates indicated an average of 0.5 and 1 SNPs per kb. A study by Van et al. (2004) of nine Korean soybean cultivars reported that SNPs in coding and non-coding regions of 110 genes occurred at a frequency of 1 per 3,260 bp ($\theta = 0.00011$) and 1 per 278 bp ($\theta = 0.00128$), respectively. They also investigated sequence diversity among 15 mapping parents of US and Korean

varieties based on 110 soybean EST sequences (Van et al. 2005). These soybeans have SNPs at a frequency of 1 per 2,038 bp in 16,302 bp of coding sequence and 1 per 191 bp in 16,960 bp of non-coding sequence. This corresponds to a nucleotide diversity (θ) of 0.00017 and 0.00186, respectively. Previous studies showed that nucleotide diversity in soybean is lower than in the autogamous species *Arabidopsis thaliana* and *Z. mays*. Hyten et al. (2006) proposed that low SNP frequency in soybean is because of the historical genetic bottleneck (by domestication, genetic drift and modern breeding) and low genetic diversity of the wild progenitor *G. soja*. The SNP identification in 102 genes revealed that modern cultivars have retained 72 % of the sequence diversity present in the Asian landraces but lost 79 % of rare alleles (frequency < 0.10) found in the Asian landraces. The lost diversity was mostly due to the small number of Asian introductions and not artificial selection. In spite of the low genetic diversity in soybean, intensive efforts in SNP discovery led to the construction of a high-density soybean genetic map joined with different maps from four mapping populations (Choi et al. 2007). For eight mapping parents including one wild soybean, aligning the sequences of 4240 genes resulted in the discovery of 5,551 SNPs at a frequency of 2.66 SNPs/ kb in introns and 2.04 SNPs/ kb in exons. Therefore, the SNP frequency in aligned soybean sequences is generally accepted to be 1 SNP per kb.

There are two recent reports on soybean sequence diversity in the soybean cyst nematode resistance gene (*GmHs1^{pro-1}*) and the determinate growth habit gene (*GmTfl1*) (Yuan et al. 2008; Tian et al. 2010). In about 1.5 kb of *GmHs1^{pro-1}*, 14 SNPs were observed in a collection including 23 wild soybeans, eight landrace species, and 13 cultivated soybean varieties, corresponding to sequence diversities of $\theta = 0.00218$ and $\pi = 0.00168$ (Yuan et al. 2008). A higher diversity of modern soybean varieties and a lower diversity of wild soybeans than landraces indicated that a narrow level of diversity for the soybean cyst nematode resistance gene was imposed by intensive selection during modern plant breeding rather than by domestication. The determinate growth habit is an agronomically important trait associated with domestication in soybean. The gene (*Dt1*) that controls determinate or indeterminate growth habit is a homolog (designated as *GmTfl1*) of the *Arabidopsis TERMINAL FLOWER1*, which is a regulatory gene that encodes a signaling protein in the shoot meristem (Liu et al. 2010). The allelic variation at the *GmTfl1* locus and the genetic diversity of a minicore collection of Chinese soybean landraces indicated that human selection for determinacy took place during the early stages of landrace radiation (Tian et al. 2010).

Due to the rapid development of NGS technology, many sequencing platforms are available, such as the GS-20 or GS-FLX from Roche/454 Life Sciences (Margulies et al. 2005), the Solexa 1G sequencer from Illumina (Bennett et al. 2005), and the SOLiD system from Applied Biosystems (<http://solid.appliedbiosystem.com>) (Parameswaran et al. 2007; Schuster 2008). Whole genome sequencing of crop species was problematic because complex genome structure and large genome sizes by high levels of the repetitive and transposable element DNA and ploidy are common (Feuillet et al. 2011). The current NGS technologies overcame these difficulties and

whole genome sequences of many economically important and polyploid crops became available, including wheat and sugarcane (*Saccharum* spp.), potato (*Solanum tuberosum*), cotton (*Gossypium hirsutum*) and banana (*Musa* spp.) (Feuillet et al. 2011; Van et al. 2011). These have been updated to produce longer read lengths and greater numbers of sequence reads. Currently, several companies are working to introduce the new technology called third generation sequencing (TGS) (Rusk 2009; Barabaschi et al. 2012). Helicos Biosciences and Pacific Biosciences developed a single molecule sequencer (Harris et al. 2008; Eid et al. 2009). The Oxford Nanopore is a new sequencing platform designed to avoid amplification, which detects a direct electrical signal (Clarke et al. 2009). Iqbal and Bashir (2011) described the fundamental principles of TGS and its applications using the Nanopore sequencer. Now, Illumina developed HiSeq 2000 as running platform at 2×150 bp. This new sequencer produced 1.13 Tb of output per run and 81 Gb daily (<http://www.illumina.com/portfolio>).

Genome resequencing by NGS technology allows us to conduct a population genetic study at the whole genome level and not in parts of targeted genomic regions. Genome resequencing of undomesticated soybean (*G. soja*) was performed (Kim et al. 2010) after the release of the draft genome sequence of cultivated soybean (*G. max*) (Schmutz et al. 2010). Using the *G. max* (var. Williams 82) genome sequence (937.5 Mb excluding gaps) as a reference, a 915.4 Mb genomic sequence of *G. soja* (var. IT182932) was determined, covering 97.65 % of the *G. max* genome sequence. The sequence difference between *G. max* (Williams 82) and *G. soja* (IT182932) was 35.2 Mb (3.76 % of 937.5 Mb), consisting of 2.5 Mb (0.267 %) of substituted bases, 406 kb (0.043 %) of indel bases, and 32.3 Mb (3.45 %) of large deleted sequences in *G. soja*. The SNPs and indels in precisely aligned areas differed by 0.31 % between Williams 82 and IT182932, and the SNP frequency was 2.67 SNPs per 1 kb. Recently, the whole genomes of 17 wild and 14 cultivated soybeans were resequenced (Lam et al. 2010). Whole genome SNP analysis revealed nucleotide diversity (θ) of 0.00189 and 0.00297 in cultivated and wild soybeans, respectively. The lower level of genetic diversity in cultivated soybeans indicated the occurrence of a bottleneck in the gene pool during domestication and under human selection. Wild-specific alleles (35 %) were more abundant than cultivated-specific alleles (5 %). In contrast to expectations, the low-frequency alleles were found to be less abundant among wild soybeans versus cultivated soybeans. This result suggests that cultivated soybean populations had expanded after domestication but the wild soybean habitat area had been reduced. In addition, genome-wide patterns of nucleotide diversity were investigated to identify conserved genomic segments shared by both wild and cultivated soybeans and genomic regions with extreme patterns of diversity and differentiation.

19.5 Model of Soybean Domestication History

It is known that cultivated soybean (*G. max*) was domesticated from its wild relative (*G. soja*) 6000–9000 years ago in China (Carter et al. 2004). Southern China, the Yellow River Valley in central China, north-eastern China, and several other regions

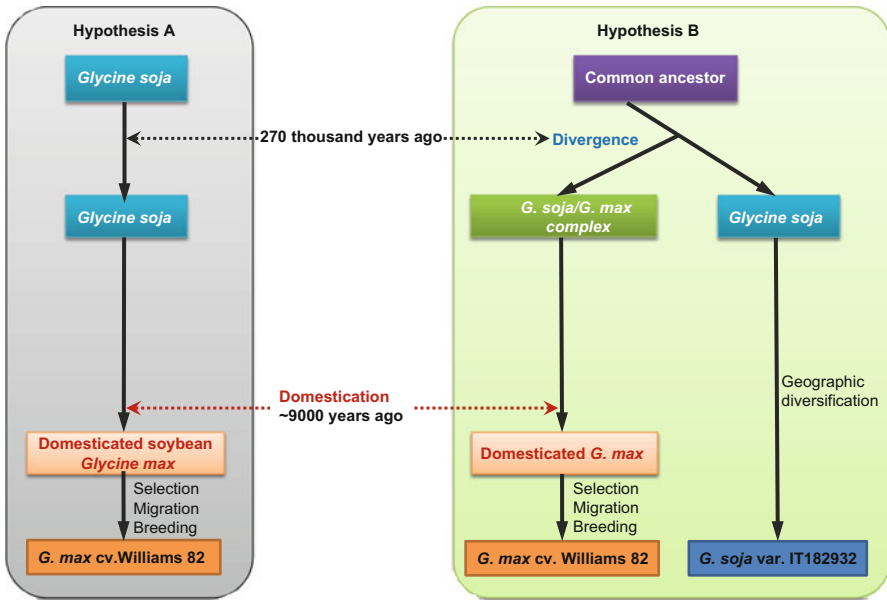


Fig. 19.1 Models of soybean domestication history. It is widely accepted that *G. max* was domesticated from its wild relative *G. soja*, ca. 6000–9000 years ago. Our estimations suggested that the *G. soja/G. max* complex diverged from a common ancestor of the two *Glycine* species ca. 270,000 years ago. Thus, divergence between *G. soja* and *G. max* predated domestication, and cultivated soybean was proposed to be domesticated from a pre-existing *G. max/G. soja* complex

(e.g., Korea and Japan) have been identified as candidate regions where soybean could have been domesticated (Carter et al. 2004). Archeological evidence suggested that both *G. soja* and *G. max* were cultivated by humans. In Chinese literature from the Shang dynasty, soybean was cultivated from 1700–1100 BC (Wilson 2008). Clearly, soybean has been cultivated much longer than previously expected. It is commonly accepted that the current cultivated soybean was domesticated from *G. soja*.

However, Kim et al. (2010) suggested that soybean was domesticated from the *G. soja/G. max* complex and diverged from a common ancestor of these two *Glycine* species, based on a calculated divergence time. Using the number of synonymous substitutions between orthologous genes to estimate the neutral genetic drift distance between these two species, it was possible to estimate the theoretical divergence time between the genomes of IT182932 (*G. soja*) and Williams 82 (*G. max*). The result indicated that *G. soja* and *G. max* diverged 0.267 ± 0.03 million years ago (Kim et al. 2010). This suggests that the divergence between IT182932 and Williams 82 predated soybean domestication (Fig. 19.1). However, archaeological evidence suggested that the domestication of soybean occurred less than 270,000 years ago and it is widely accepted that *G. max* is essentially a domesticated form of *G. soja* (Fig. 19.1). There are many possible reasons for this discrepancy. Kim et al. (2010) used only two single genomes for comparisons. The *G. max* genome might have incorporated genetic

material from multiple species involved in the domestication process, resulting in the relatively larger genetic distance. Since *G. soja* and *G. max* were present at the same time period (Lee and Park 2006), there may have been an undomesticated population of *G. max* somewhere in East Asia that became the direct ancestor of the domesticated population. The domestication process itself was influenced by human selection, so the divergence time between *G. soja* and *G. max* might be overestimated. Thus, Kim et al. (2010) suggested that the domestication history of soybean is more complicated than previously assumed. Additional studies, such as the sequencing and comparison of more soybean genomes, are needed to establish the origin of domesticated *G. max* and the processes underlying domestication.

19.6 Conclusions

Human selection is the most influential factor underlying soybean domestication, and many useful genes such as protein content and disease resistance may be lost. Both archeological and molecular evidence suggested that gathering and cultivation were practiced prior to domestication. Since domestication represents a rapid change in evolution and leads to decreasing genetic diversity, studying wild and domesticated populations will be helpful. To overcome the narrow genetic background of the cultivated soybean, comparing sequence diversity around DRTs between cultivated soybean and its progenitor would give a new view of the process of domestication.

Genome sequencing enabled a study of the evolution of DRTs (Vaughan et al. 2007). The rapid advances in massively parallel sequencing technologies will allow complete genome comparisons among cultivars and progenitors of crop plants and will provide information on genes that are absent or present in cultivated soybean (*G. max*) (Kim et al. 2010). In this review, we have described the genome sequencing of wild soybean and suggested that soybean was domesticated from the *G. soja*/*G. max* complex that diverged from a common ancestor of these two *Glycine* species. The genome sequences of wild species may provide key information about the genetic elements involved in speciation and domestication.

A genome-wide comparison of wild and domesticated soybeans allowed us to characterize unique features and differences within the genomes, and to derive useful information about soybean domestication. The hundreds of genes with putative functional differences between *G. soja* and *G. max* could be useful candidate genes for reverse genetic studies of crop domestication. Extensive studies by resequencing or *de novo* sequencing of soybean at a population level are needed for understanding the genomic impacts on soybean domestication history (Henry 2012; Van et al. 2013). Also, these studies will potentially enable us to isolate genes controlling DRTs and to determine which genes were targeted by selection during soybean domestication.

Acknowledgment This research was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ008117) of the Rural Development Administration, Republic of Korea.

References

- Andersen JR, Lubberstedt T (2003) Functional markers in plants. *Trend Plant Sci* 8:554–560
- Bailey MA, Mian MAR, Carter TE et al (1997) Pod dehiscence of soybean: identification of quantitative trait loci. *J Hered* 88:152–154
- Barabaschi D, Guerra D, Lacrima K et al (2012) Emerging knowledge from genome sequencing of crop species. *Mol Biotechnol* 50:250–266
- Bennett ST, Barnes C, Cox A et al (2005) Toward the 1,000 dollars human genome. *Pharmacogenomics* 6:373–382
- Boerma HR, Specht JE (2004) Soybeans: improvement, production and uses. Am Soc of Agro, Madison
- Buckler ES, Thornsberry JM, Kresovich S (2001) Molecular diversity, structure and domestication of grasses. *Genet Res* 77:213–218
- Burke JM, Tang S, Knapp SJ, Rieseberg LH (2002) Genetic analysis of sunflower domestication. *Genetics* 161:1257–1267
- Cai HW, Morishima H (2000) Genomic regions affecting seed shattering and seed dormancy in rice. *Theor Appl Genet* 100:840–846
- Cai HW, Morishima H (2002) QTL clusters reflect character associations in wild and cultivated rice. *Theor Appl Genet* 104:1217–1228
- Carter TE Jr, Nelson R, Sneller CH, Cui Z (2004) Genetic diversity in soybean. In: Boerma HR, Specht JE (eds) Soybeans: improvement, production and uses. Am Soc of Agro, Madison, pp 303–416
- Choi I-Y, Hyten DL, Matukumalli LK et al (2007) A soybean transcript map: gene distribution, haplotype and single-nucleotide polymorphism analysis. *Genetics* 176:685–696
- Clarke J, Wu H-C, Jayasinghe L et al (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnol* 4:265–270
- Concibido VC, La Vallee B, McIaird P et al (2003) Introgression of a quantitative trait locus for yield from *Glycine soja* into commercial soybean cultivars. *Theor Appl Genet* 106:575–582
- Crawford G (2005) East Asian plant domestication. In: Stark MT (ed) *Archaeology of asia*. Blackwell Publishing, Oxford, pp 77–95
- Crawford G, Lee G-A (2003) Agricultural origins in the Korean Peninsula. *Antiquity* 77:87–95
- Crawford GW, Underhill AP, Zhao J et al (2005) Late neolithic plant remains from northern China: preliminary results from Liangchengzhen, Shandong. *Curr Anthropol* 46:309–317
- Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* 418:700–707
- Doebley JF (1989) Isozymic evidence and the evolution of crop plants. In: Soltis D, Soltis P (eds) *Isozymes in plant biology*. Dioscorides Press, Portland, pp 165–191
- Doebley J, Stec A (1993) Inheritance of the morphological differences between maize and Teosinte—comparison of results for two F₂ populations. *Genetics* 134:559–570
- Doebley J, Stec A, Wendel J, Edwards M (1990) Genetic and morphological analysis of a maize Teosinte F₂ population—implications for the origin of maize. *Proc Natl Acad Sci USA* 87:9888–9892
- Doebley JF, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127:1309–1321
- Doganlar S, Frary A, Daunay MC et al (2002) Conservation of gene function in the *Solanaceae* as revealed by comparative mapping of domestication traits in eggplant. *Genetics* 161:1713–1726
- Dong YS, Zhao LM, Liu B et al (2004) The genetic diversity of cultivated soybean grown in China. *Theor Appl Genet* 108:931–936
- Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138
- Feuillet C, Leach JE, Rogers J et al (2011) Crop genome sequencing: lessons and rationales. *Trend Plant Sci* 16:77–88

- Frery A, Nesbitt TC, Frery A et al (2000) *fw2.2*: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289:85–88
- Fuller DQ (2002) Fifty years of archaeobotanical studies in India: laying a solid foundation. In: Settar S, Korisettar R (eds) *Indian archaeology in retrospect: archaeology and interactive disciplines*, vol III. Manohar, Delhi, pp 247–263
- Fuller DQ (2007) Contrasting patterns in crop domestication and domestication rates: recent archaeobotanical insights from the old world. *Ann Bot* 100:903–924
- Funatsuki H, Ishimoto M, Tsuji H et al (2006) Simple sequence repeat markers linked to a major QTL controlling pod shattering in soybean. *Plant Breed* 125:195–197
- Grandillo S, Tanksley SD (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicon pimpinellifolium*. *Theor Appl Genet* 92:935–951
- Gross BL, Olsen KM (2010) Genetic perspectives on crop domestication. *Trend Plant Sci* 15:529–537
- Gunter C (2008) Plant genetics rice stands up. *Nat Rev Genet* 9:816–816
- Guo J, Wang Y, Song C et al (2010) A single origin and moderate bottleneck during domestication of soybean (*Glycine max*): implications from microsatellites and nucleotide sequences. *Ann Bot* 106:505–514
- Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. *Euphytica* 156:1–13
- Harris DR (1998) The origins of agriculture in southwest Asia. *Rev Archaeol* 19:5–11
- Harris TD, Buzby PR, Babcock H et al (2008) Single-molecule DNA sequencing of a viral genome. *Science* 320:106–109
- Henry RJ (2012) Next-generation sequencing for understanding and accelerating crop domestication. *Brief Func Genom* 11:51–56
- Hymowitz T (1970) On the domestication of the soybean. *Eco Bot* 24:408–421
- Hyten DL, Song Q, Zhu Y et al (2006) Impacts of genetic bottlenecks on soybean genome diversity. *Proc Natl Acad Sci USA* 103:16666–16671
- Iqbal SM, Bashir R (2011) *Nanopores: sensing and fundamental biological interactions*. Springer, New York
- Jackson SA, Iwata A, Lee SH et al (2011) Sequencing crop genomes: approaches and applications. *New Phytol* 191:915–925
- Jin J, Huang W, Gao JP et al (2008) Genetic control of rice plant architecture under domestication. *Nature Genet* 40:1365–1369
- Jun T-H, Van K, Kim MY et al (2011) Uncovering signatures of selection in the soybean genome using SSR diversity near QTLs of agronomic importance. *Genes Genom* 33:391–397
- Kang S-T, Kwak M, Kim H-K et al (2009) Population-specific QTLs and their different epistatic interactions for pod dehiscence in soybean [*Glycine max* (L.) Merr.]. *Euphytica* 166:15–24
- Keim P, Diers BW, Olson TC, Shoemaker RC (1990a) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. *Genetics* 126:735–742
- Keim P, Diers BW, Shoemaker RC (1990b) Genetic analysis of soybean hard seededness with molecular markers. *Theor Appl Genet* 79:465–469
- Kim MY, Lee S, Van K et al (2010) Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. *Proc Natl Acad Sci USA* 107:22032–22037
- Kim MY, Van K, Kang YJ et al (2012) Tracing soybean domestication history: from nucleotide to genome. *Breed Sci* 61:445–452
- Koinange EMK, Singh SP, Gepts P (1996) Genetic control of the domestication syndrome in common bean. *Crop Sci* 36:1037–1045
- Kovach MJ, Sweeney MT, McCouch SR (2007) New insights into the history of rice domestication. *Trend Genet* 23:578–587
- Lam H-M, Xu X, Liu X et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genet* 42:1053–1059

- Lauter N, Doebley J (2002) Genetic variation for phenotypically invariant traits detected in teosinte: implications for the evolution of novel forms. *Genetics* 160:333–342
- Lee G-A (2011) The transition from foraging to farming in prehistoric Korea. *Curr Anthropol* 52:S307–S329
- Lee Y-H, Park T-S (2006) Origin of legumes cultivation in Korean Peninsula by viewpoint of excavated grain remains and genetic diversity of legumes. *Kor Agri Hist Assoc* 5:1–31 (in Korean)
- Lee G-A, Crawford GW, Liu L, Chen X (2007) Plants and people from the early neolithic to shang periods in North China. *Proc Natl Acad Sci USA* 104:1087–1092
- Lee JD, Yu JK, Hwang YH et al (2008) Genetic diversity of wild soybean (*Glycine soja* Sieb. and Zucc.) accessions from South Korea and other countries. *Crop Sci* 48:606–616
- Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering. *Science* 311:1936–1939
- Li DD, Pfeiffer TW, Cornelius PL (2008) Soybean QTL for yield and yield components associated with *Glycine soja* alleles. *Crop Sci* 48:571–581
- Li YH, Guan RX, Liu ZX et al (2008) Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China. *Theor Appl Genet* 117:857–871
- Li YH, Li W, Zhang C et al (2010) Genetic diversity in domesticated soybean (*Glycine max*) and its wild progenitor (*Glycine soja*) for simple sequence repeat and single-nucleotide polymorphism loci. *New Phytol* 188:242–253
- Liu B, Fujita T, Yan Z-H et al (2007) QTL mapping of domestication-related traits in soybean (*Glycine max*). *Ann Bot* 100:1027–1038
- Liu B, Watanabe S, Uchiyama T et al (2010) The soybean stem growth habit gene *Dt1* is an ortholog of *Arabidopsis TERMINAL FLOWER1*. *Plant Physiol* 153:198–210
- Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380
- Nagai YS, Sobrizal, Sanchez PL et al (2002) *Sh3*, a gene for seed shattering, commonly found in wild rices. *Rice Genet Newsl* 19:74–75
- Nielsen R (2005) Molecular signatures of natural selection. *Annu Rev Genet* 39:197–218
- Parameswaran P, Jalili R, Tao L et al (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 35:e130
- Peng JH, Ronin Y, Fahima T et al (2003) Domestication quantitative trait loci in *Triticum dicoccoides*, the progenitor of wheat. *Proc Natl Acad Sci USA* 100:2489–2494
- Poncet V, Martel E, Allouis S et al (2002) Comparative analysis of QTLs affecting domestication traits between two domesticated x wild pearl millet (*Pennisetum glaucum* L., Poaceae) crosses. *Theor Appl Genet* 104:965–975
- Purugganan MD, Fuller DQ (2009) The nature of selection during plant domestication. *Nature* 457:843–848
- Ross-Ibarra J (2005) Quantitative trait loci and the study of plant domestication. *Genetica* 123:197–204
- Rusk N (2009) Cheap third-generation sequencing. *Nat Methods* 6:244–245
- Sang T, Ge S (2007) The puzzle of rice domestication. *J Integr Plant Biol* 49:760–768
- Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the paleopolyploid soybean. *Nature* 463:178–183
- Schulman AH (2007) Molecular markers to assess genetic diversity. *Euphytica* 158:313–321
- Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Methods* 5:16–18
- Sobrizal, Ikeda K, Sanchez PL, Yoshimura A (1999) RFLP mapping of a seed shattering gene on chromosome 4 in rice. *Rice Genet Newsl* 16:74–75
- Stupar RM (2010) Into the wild: the soybean genome meets its undomesticated relative. *Proc Natl Acad Sci USA* 107:21947–21948
- Tang H, Sezen U, Paterson AH (2010) Domestication and plant genomes. *Curr Opin Plant Biol* 13:160–166
- Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol Biol Evol* 21:1214–1225

- Tian Z, Wang X, Lee R et al (2010) Artificial selection for determinate growth habit in soybean. *Proc Natl Acad Sci USA* 107:8563–8568
- Van K, Hwang E-Y, Kim MY et al (2004) Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs. *Euphytica* 139:147–157
- Van K, Hwang E-Y, Kim MY et al (2005) Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. *J Hered* 96:529–535
- Van K, Kim D, Cai CM et al (2008) Sequence level analysis of recently duplicated regions in soybean [*Glycine max* (L.) Merr.] genome. *DNA Res* 15:93–102
- Van K, Kim DH, Shin JH, Lee S-H (2011) Genomics of plant genetic resources: past, present and future. *Plant Genet Resour* 9:155–158
- Van K, Rastogi K, Kim K-H, Lee S-H (2013) Next-generation sequencing technology for crop improvement. *SABRAO J Breed Genet* 45:84–99
- Vaughan DA, Balazs E, Heslop-Harrison JS (2007) From crop domestication to super-domestication. *Ann Bot* 100:893–901
- Vielle-Calzada JP, Martinez delaVO, Hernandez-Guzman G et al (2009) The Palomero genome suggests metal effects on domestication. *Science* 326:1078
- Wilson RF (2008) Soybean: market driven research needs. In: Stacey G (ed) *Genetics and genomics of soybean*. Plant genetics and genomics, vol 2. Springer, pp 3–15
- Xiong LX, Liu KD, Dai XK et al (1999) Identification of genetic factors controlling domestication-related traits of rice using an F-2 population of a cross between *Oryza sativa* and *O. rufipogon*. *Theor Appl Genet* 98:243–251
- Xu DH, Gai JY (2003) Genetic diversity of wild and cultivated soybeans growing in China revealed by RAPD analysis. *Plant Breed* 122:503–506
- Xu DH, Abe J, Gai JY, Shimamoto Y (2002) Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. *Theor Appl Genet* 105:645–653
- Yuan C, Zhou G, Li Y, Wang K, Wang Z, Li X, Chang R, Qiu L (2008) Cloning and sequence diversity analysis of *GmHs1^{pro-1}* in Chinese domesticated and wild soybeans. *Mol Breed* 22:593–602
- Zhu YL, Song QJ, Hyten DL et al (2003) Single-nucleotide polymorphisms in soybean. *Genetics* 163:1123–1134

Chapter 20

Genomics of Origin, Domestication and Evolution of *Phaseolus vulgaris*

Elisa Bellucci, Elena Bitocchi, Domenico Rau, Monica Rodriguez, Eleonora Biagetti, Alessandro Giardini, Giovanna Attene, Laura Nanni and Roberto Papa

Contents

20.1 Introduction	484
20.2 Origin of the Common Bean	485
20.3 Domestication of <i>P. vulgaris</i>	489
20.4 Diffusion and Evolution of <i>P. vulgaris</i> Out of the American Centers of Origin	492
20.4.1 The Contribution of the Mesoamerican and Andean Gene Pools to Bean Germplasm Collections from Different Parts of the World	492
20.4.2 The Diversity of the 'out-of-America' Germplasm, and Their Divergence from the American Source Population	494
20.4.3 Introgression Between the Mesoamerican and Andean Gene Pools	496
20.5 Genomic Tools and Germplasm Collections	497
20.6 Conclusions	502
References	502

Abstract The role of genetic diversity is crucial for future improvements to meet societal demand for food security under a climate change scenario. From this perspective, it is thus crucial to understand the structure and evolution of crop species and their wild relatives. The common bean (*Phaseolus vulgaris* L.) is the world's most important food legume for direct use, and the demand for this crop can be expected to increase based on the current trends in population growth and bean consumption. The wild *P. vulgaris* has a Mesoamerican origin, and since its expansion, it has become distributed from northern Mexico to north-western Argentina, which has led to the formation of two major gene pools in these geographical regions. Domestication took place after the formation of these gene pools, and their structure is

R. Papa (✉) · E. Bellucci · E. Bitocchi · E. Biagetti · A. Giardini · L. Nanni
Dipartimento di Scienze Agrarie, Alimentari ed Ambientali, Università Politecnica delle Marche,
via Breccie Bianche, 60131 Ancona, Italy
e-mail: r.papa@univpm.it

D. Rau · M. Rodriguez · G. Attene
Dipartimento di Agraria, Università degli Studi di Sassari, 07100 Sassari, Italy

R. Papa
Consiglio per la Ricerca e Sperimentazione in Agricoltura, Cereal Research Centre (CRA-CER),
S.S. 16, Km 675, 71122 Foggia, Italy

still clearly evident in both the wild and the domesticated forms. This evolutionary scenario renders *P. vulgaris* almost unique among crops, and therefore particularly useful to investigate crop domestication, as this process can be studied in the same species as a replicated experiment (i.e., in Mesoamerica and in the Andes). The present review offers an overview of the current knowledge on the evolutionary history of *P. vulgaris* L. including speciation, domestication, diversification, and crop expansion outside its centers of domestication in Mesoamerica and in the Andes. Within this context, we also present a description of the available genomic tools and the germplasm collections that are at present available for genetic studies on the common bean, while showing their potential for improvements to the productivity and quality of this crop.

Keywords Common bean · Crop evolution · Genetic resources · Molecular diversity · Pre-breeding

20.1 Introduction

Legumes represent an important component of agricultural food crops and they have a crucial role in both farming systems and the human diet, especially in developing countries. Globally, legumes complement cereal crops as a source of protein and minerals, with a harvested area of about one-tenth that collectively under cereals (Akibode and Maredia 2011).

The important role of food legumes in the farming systems is underlined by the demonstration that over the past 15 years overall legume production has increased at a greater rate than the growth rate of the world population (Akibode and Maredia 2011). Among pulse crops, *Phaseolus* is a large and diverse genus that comprises about 70 species from Central and North America (Freytag and Debouck 2002), five of which have been domesticated (*P. vulgaris*, *P. dumosus*, *P. coccineus*, *P. acutifolius*, *P. lunatus*), and with a few additional species that show signs of incipient domestication (Delgado-Salinas et al. 2006).

The common bean (*Phaseolus vulgaris* L.) is the world's most important food legume for direct use, with a production of about 12 million metric tons per year. The leading countries in this production are Latin America and sub-Saharan Africa, where three-quarters of this crop is grown (<http://faostat.fao.org/>; Akibode and Maredia 2011).

Considering the current trends in population growth and bean consumption, the demand for *P. vulgaris* can be expected to increase (CIAT report 2001, <http://webapp.ciat.cgiar.org/ciatinfocus/beans.htm>; Akibode and Maredia 2011), and compelling questions about this species must be addressed in the future. Bean productivity, food quality, and resistance to biotic and abiotic factors, among others, would realistically be the aim for future investigations to meet the challenges posed by climate change and the fast increasing demand for food.

It has been shown that an acceleration of the rate of crop improvement can be achieved by taking advantage of high-throughput genomic technologies that are having significant effects on the management of gene banks and on the way germplasm collections are exploited (Tuberosa et al. 2011). Different tools and sources of genomic information on the bean genome are nowadays available to investigate the diversity present in this species, including molecular linkage maps, expressed sequence tags (EST) collections, bacterial artificial chromosome libraries, a physical map, and soon, a whole-genome sequence (McClean et al. 2008, 2013; Gepts et al. 2008; <http://www.phytozome.net/commonbean.php>). The advent of next-generation sequencing has revolutionized genomic and transcriptomic approaches to biology (Gupta et al. 2008; Mardis 2008). These new sequencing tools are also valuable for single nucleotide polymorphism (SNP) discovery and the detection of genetic markers in populations, which in turn, can be exploited in different studies (Davey et al. 2011).

Despite there still being limitations to contemporary common bean breeding, genomics-assisted techniques have been widely exploited in this species, and have enhanced the effectiveness of breeding programs and responses to selection (Beaver and Osorno 2009; Tuberosa et al. 2011). Marker-assisted selection is often routinely used for traits controlled by major loci, although marker-assisted selection for complex quantitative traits still remains a challenging task in breeding programs (Miklas et al. 2006; Tuberosa et al. 2011). Genome-wide association mapping is an approach that is being increasingly adopted to dissect out the genetic basis of target traits, and when it is applied to wild populations, it has substantial benefits for conservation genetics and ecology (Allendorf et al. 2010; Galeano et al. 2012). Concurrently, diversity analyses and evolution of the species can be understood by investigating domestication, local adaptation, genetic drift, and gene flow through novel genomic techniques (Davey et al. 2011; van Heerwaarden et al. 2011).

The lack of a whole genome sequence for the common bean has been a major limitation for such an important crop species. The forthcoming availability for the scientific community of the biotechnology tools that are available for other crops will enhance the competitiveness of this species. Realistically, the sequence will provide powerful tools to improve agronomic and nutritional traits, which is particularly important to maintain and improve the nutritional status of poor individuals. Future genomic studies will contribute to the gaining of insights into this important crop, such as comparative gene discovery in legumes, fine-mapping and candidate gene identification, and the identification of *Phaseoleae* domestication and adaptation genes.

20.2 Origin of the Common Bean

According to the geographical distribution of most of the species belonging to the *Phaseolus* genus, these are considered to be of Mesoamerican origin (Freytag and Deboucq 1996, 2002; Delgado-Salinas et al. 1999, 2006). Delgado-Salinas et al. (2006) analyzed internal transcribed spacers (ITS) of the ribosomal DNA and the chloroplast *trnK* locus, and they showed that the *Phaseolus* crown clade is no older

than *ca.* 4–6 My. The present-day form of Mexico was apparent by the Late Miocene (5 My ago), with a final major event of subduction volcanism that resulted in the modern Trans-Mexican Volcanic Belt. This strongly suggests that *Phaseolus* diversification took place during and after this major tectonic activity (Delgado-Salinas et al. 2006), and thus evolved well after the period when the land bridge connecting Mesoamerica and South America was formed, which was *ca.* 7 My ago (Coates et al. 2004). Delgado-Salinas et al. (2006) detected eight principal crown clades within *Phaseolus*, with the *vulgaris* group as the oldest, at *ca.* 4 My. This group includes four of the five domesticated species of the genus (*P. vulgaris*, *P. dumosus*, *P. coccineus*, *P. acutifolius*). The closest relatives to *P. vulgaris* are the Mesoamerican species *P. dumosus* and *P. coccineus*, and these three species together are partially intercrossable. The other domesticated species (*P. lunatus*, *P. acutifolius*) are more distantly related. On the basis of sequence data of the α -amylase inhibitor gene, *P. vulgaris* diverged from *P. dumosus* and *P. coccineus* *ca.* 2 My ago (Gepts et al. 1999).

Among the five domesticated *Phaseolus* species, *P. vulgaris* is the most important economically, as it is the main grain legume for direct human consumption. It is a rich source of protein, vitamins, minerals and fiber, especially in less-developed countries (http://www.fao.org/index_en.htm, 2010; Broughton et al. 2003). *P. vulgaris* is a true autogamous diploid species, with 22 chromosomes and a haploid genome size that is estimated to be between 587 Mbp and 637 Mbp (Arumuganathan and Earle 1991; Bennett and Leitch 1995, 2010).

Wild *P. vulgaris* is widely distributed from northern Mexico to north-western Argentina (Toro et al. 1990), and it is characterized by two major eco-geographical gene pools: those of Mesoamerica and the Andes. These two gene pools show parallel wild and domesticated geographical structures, as shown by several studies based on different datasets, including plant morphology (Singh et al. 1991b), seed proteins (Gepts et al. 1986; Gepts and Bliss 1985), allozymes (Koenig and Gepts 1989), restriction fragment length polymorphism (Becerra-Velásquez and Gepts 1994), random amplified polymorphic DNA (RAPD, Freyre et al. 1996), amplified fragment length polymorphism (AFLP; Papa and Gepts 2003; Rossi et al. 2009), and simple sequence repeats (microsatellites, SSRs; Kwak and Gepts 2009).

In the 1980's, a wild *P. vulgaris* population was discovered in northern Peru and Ecuador (Debouck et al. 1993). Kami et al. (1995) analyzed a portion of the gene that codes for the seed protein phaseolin, and they identified a new phaseolin type (type I) for this population from northern Peru–Ecuador that was not present in the other gene pools, thus indicating that this population is a new distinct wild gene pool. The type I phaseolin gene does not carry the tandem direct repeats that are present in Mesoamerican and Andean accessions. With the consideration that duplications that generate tandem direct repeats are more likely than deletions that specifically eliminate a member of a tandem direct repeat, Kami et al. (1995) suggested that type I phaseolin is ancestral to the other phaseolin sequences of *P. vulgaris*. This arises because duplications can occur in many locations along a sequence, whereas deletions can occur only at the site of the tandem direct repeats. Thus, the most credited hypothesis on the origin of the common bean was that from the core area of the

western slopes of the Andes in northern Peru and Ecuador, from where the wild bean was dispersed northwards (Colombia, Central America and Mexico) and southwards (southern Peru, Bolivia and Argentina), which resulted in the Mesoamerican and Andean gene pools, respectively (Kami et al. 1995). However, recently, this hypothesis has been called into question by different studies (Rossi et al. 2009; Nanni et al. 2011; Bitocchi et al. 2012, 2013; Desiderio et al. 2013). In particular, Bitocchi et al. (2012) clearly indicated a Mesoamerican origin of the common bean by investigating the nucleotide diversity at five different gene fragments on a wide sample of wild *P. vulgaris* that is representative of its geographical distribution.

The first evidence towards this statement was the occurrence of a bottleneck prior to domestication for the Andean gene pool. This is supported by the higher genetic diversity detected for the Mesoamerican gene pool, as compared to the Andean gene pool, which resulted in a 90 % loss of diversity for the Andean gene pool (Bitocchi et al. 2012). This trend had already been reported in earlier (Freyre et al. 1996; Koenig and Gepts 1989) and more recent studies (Kwak and Gepts 2009; Rossi et al. 2009; Nanni et al. 2011; Bitocchi et al. 2013; Desiderio et al. 2013). However, the genetic diversity reduction using sequence data was about two-fold, 13-fold and three-fold higher than those in a comparable sample of *P. vulgaris* genotypes using AFLP data (45 %; Rossi et al. 2009), SSR data (7 %; Kwak and Gepts 2009) and chloroplast (cp)SSR data (26 %; Desiderio et al. 2013), respectively. This is clear evidence of the crucial role of marker mutation rates for describing the diversity of plant populations (Thuillet et al. 2005). In particular, the loss of diversity detected with cpSSRs is intermediate between the SSRs and AFLPs, as is their mutation rate (Provan et al. 1999; Marshall et al. 2002). Indeed, as showed in several studies (Glémin and Bataillon 2009, Rossi et al. 2009; Nanni et al. 2011; Bitocchi et al. 2012, 2013; Desiderio et al. 2013), in populations that have experienced a bottleneck, the differences in loss of diversity estimates using different markers are related to their different mutation rates: in markers characterized by high mutation rates, such as SSRs, the recovery of the diversity lost after a bottleneck is faster than for markers with lower mutation rates, such as sequence data.

The second novel outcome of the analysis carried out using sequence data was the population structure identified in Mesoamerica. Indeed, before the study of Bitocchi et al. (2012), even if it was known that the wild Mesoamerican gene pool was characterized by a high population structure (Papa and Gepts 2003), a clear distinction into groups had never been found, and thus Mesoamerica was usually considered as a single gene pool. The main reason for this was probably related to the nature of the markers used; indeed, previous studies did not clearly detect any population subdivisions in Mesoamerica due to hybridization and recombination between the different groups, which reduced the discriminatory power of the multilocus molecular markers used (Kwak and Gepts 2009; Rossi et al. 2009). As sequence data are less prone to these factors, Bitocchi et al. (2012) showed that the Mesoamerican accessions can indeed be split into four distinct genetic groups: B1, B2, B3 and B4 (Fig. 20.1). The B1 group was represented by accessions distributed across all of the geographical area, from the north of Mexico down to Colombia. The other three groups were composed of only Mexican accessions. The B2 group was spread

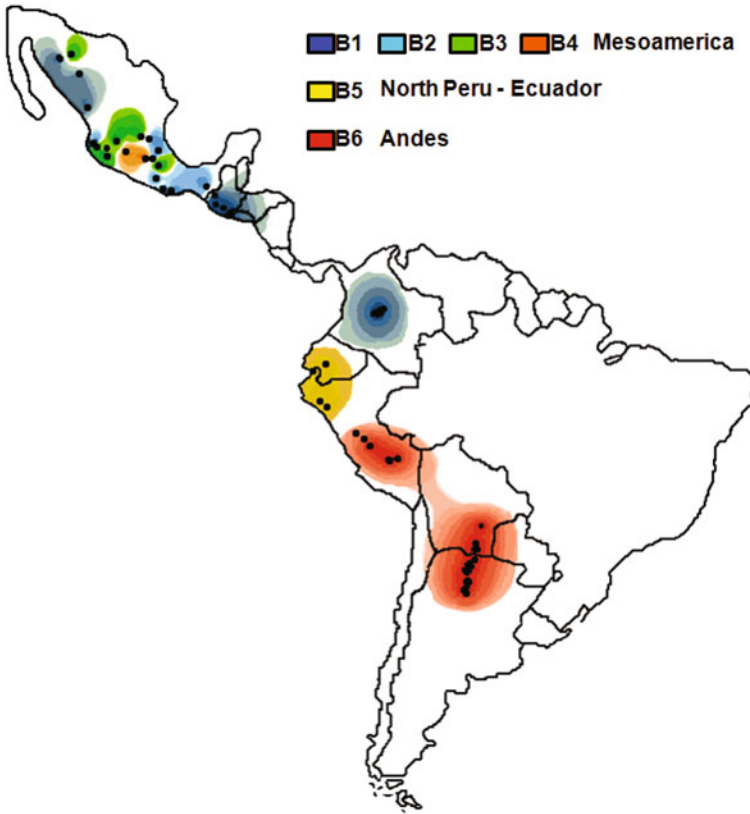


Fig. 20.1 Representation of the geographical distribution of the clusters identified by the Bayesian clustering analysis performed by Bitocchi et al. (2012). See legend for correspondence between colors and genetic clusters

from central to southern Mexico, while the B3 and B4 groups were present in a wide area of central Mexico (Fig. 20.1). Investigations into the relationships between these different groups have shown that, remarkably, there is no clear distinction between the Mesoamerican and Andean wild gene pools, while different relationships of the Mesoamerican groups with the north Peru–Ecuador and Andean gene pools were found (Bitocchi et al. 2012). In particular, the Andean wild accessions were more related to the Mesoamerican B3 accessions, and the northern Peru–Ecuador accessions to the Mesoamerican B4 accessions (Fig. 20.1). The Bitocchi et al. (2012) study shows clear evidence of a Mesoamerican origin of the common bean, which was most likely located in Mexico, which is consistent with the known distribution of most of the close relatives of *P. vulgaris*. Thus, both of the gene pools from South America originated through different migration events from the Mesoamerica populations of central Mexico. These results are strongly supported by those obtained at chloroplast DNA level on a partially overlapping sample of wild accessions

(Desiderio et al. 2013). Bitocchi et al. (2012) suggested that the wild common bean from northern Peru and Ecuador is a relict population that only represents a fraction of the genetic diversity of the ancestral population. Considering that the results of Kami et al. (1995) that indicated that phaseolin type I (PhI) is an ancestral phaseolin are relatively robust, the absence of this phaseolin type in Mesoamerica would be due to its extinction in this gene pool, or alternatively, it might still be present, but just not included in the samples analyzed in the literature.

20.3 Domestication of *P. vulgaris*

Domestication is a complex process that modifies a wild plant and makes it into a crop. In *P. vulgaris*, this involved several morphological and physiological changes, such as differences in growth habit (indeterminate vs determinate), seed dormancy (present vs not present), photoperiod sensitivity (short-day vs insensitivity), shape, color and size of the plant and its harvested parts, and the dissemination mechanisms (shattering vs non-opening pods). All of these structural and functional modifications shared among most crop species (the domestication syndrome) make them genetically different from their wild types, and confer better adaptation to different agro-ecosystems (Gepts and Papa 2002). The process of common-bean domestication has been studied in detail, and the major domestication traits have been mapped (Koinange et al. 1996). Koinange et al. (1996) performed a quantitative trait locus (QTL) analysis using a recombinant inbred population derived from a cross of wild × cultivated, and they found that the QTLs for the traits measured tended to cluster into several regions on the bean linkage map. Some of the candidate genes associated with the domestication process have been characterized (Anthony et al. 1990; Kwak et al. 2006, 2008; Repinski et al. 2012).

One of the consequences of domestication that is common to most crop species is the reduction of genetic diversity due to a founder effect (Glémin and Bataillon, 2009). In analyzing Mesoamerican and Andean wild and domesticated populations using AFLP markers, Rossi et al. (2009) observed a strong reduction in the genetic diversity due to domestication (wild vs domesticated samples) only in the Mesoamerica population ($\Delta H = 0.32$). Markers that differ substantially in their mutation rates can show very different patterns of molecular diversity, and indeed, Kwak and Gepts (2009) used SSR markers to show a lower reduction in Mesoamerica (ca. 10%).

The data from Nanni et al. (2011) from the analysis of a genomic sequence in the wild and domesticated common bean that is similar to SHATTERPROOF 1 (*PvSHP1*), the gene involved in the control of fruit shattering in *Arabidopsis thaliana*, offered the first estimates of the effects of domestication on nucleotide variation in this species, based on a relatively large and representative sample of genotypes. The loss of diversity in the domesticated accessions in the Andes was 54%; in Mesoamerica, this loss of diversity ranged from 65–69% when compared with only the wild accessions from Mexico, and with all of the Mesoamerican wild populations, respectively. These results have been confirmed more recently by the analysis of five gene fragments in 214 accessions (102 wild and 112 domesticated) of *P. vulgaris* (Bitocchi et al. 2013). Indeed, it was shown that the domestication of the common

bean in Mesoamerica induced a severe reduction (72 %) in genetic diversity, which was consistently reproduced for all of the five genes studied (range, 44–98 %). Additionally, the pattern was also confirmed in the Andes data (loss of diversity, 27 %). However, the reduction in genetic diversity was three-fold greater in Mesoamerica compared with the Andes. As proposed by Bitocchi et al. (2013), this difference can be explained as the result of the bottleneck that occurred before domestication in the Andes (Rossi et al. 2009; Bitocchi et al. 2012), which strongly impoverished the Andean wild populations, leading to the minor effects of the subsequent domestication bottleneck (i.e., sequential bottleneck). These findings show the importance of considering the evolutionary history of a crop species as a major factor that influences its current level and structure of genetic diversity.

Papa et al. (2005) showed that genes for domestication are located in regions of high divergence between wild and domesticated *P. vulgaris*. Also, the regions linked to the domestication loci have probably been less exploited by farmers and breeders, and these are the ones where the highest diversity of the wild relatives is located. Several studies have clearly indicated that the use of wild relatives can have a tremendous impact on crop improvement (Tanskley and McCouch 1997; McCouch 2004); therefore, to better exploit the genetic diversity that is present in the wild relatives of a crop, knowledge of the locations of the genes involved in the domestication syndrome and the proportion of the genome affected by domestication appears to be crucial. This knowledge of the domestication loci is indeed useful in two main ways: for identification of markers that are tightly linked to undesirable genes (e.g., shattering); and for the possibility to identify the surrounding chromosomal regions that would be most likely to harbor the highest and historically less exploited diversity of the wild germplasm. Using the approach of a genome scan for the signature of domestication, Papa et al. (2007) estimated that a large fraction of the genome of the common bean appears to be under the effects of selection during domestication (about 16 %). Molecular analysis was carried out using 2,506 AFLP markers on 14 bulks of individuals (seven bulks of wild, and seven bulks of domesticated). For the allelic frequencies of the wild and domesticated populations based on the bulk analysis, and for each marker, in both datasets, the departure from the neutral expectation was evaluated using a method based on F_{ST} , to identify loci that were putatively under selection. Moreover, AFLP markers analyzed on single genotypes were mapped on a *P. vulgaris* consensus map, and most of those that were putatively under the effects of selection due to the domestication loci were localized close to genes and QTLs that are linked to the domestication process.

The common bean was domesticated independently in Mesoamerica and in the Andes. Two independent domestication events in the Americas have been documented in several studies, and a large set of coherent data have been obtained using different approaches based on molecular markers and morphological characteristics (Gepts et al. 1986; Gepts and Bliss 1988; Koenig and Gepts 1989; Gepts and Debouk 1991; Singh et al. 1991a, b, c; Becerra Velasquez and Gepts 1994; Freyre et al. 1996; Tohme et al. 1996; Gepts 1998; Delgado-Salinas et al. 1999; Papa and Gepts 2003; Blair et al. 2006a, b; Diaz and Blair 2006; Angioi et al. 2009a; Kwak and Gepts 2009; Rossi et al. 2009; Nanni et al. 2011; Blair et al. 2012; Bitocchi et al. 2013). These two

independent domestication events, one in Mesoamerica and one in the Andes, gave origin to two major domesticated gene pools (Papa et al. 2006; Acosta-Gallegos et al. 2007; Angioi et al. 2009a). Following domestication, the domesticated gene pools of the common bean appear to have been organized into four Mesoamerican (*Durango*, *Jalisco*, *Mesoamerica*, *Guatemala*) and three Andean (*Nueva Granada*, *Peru*, *Chile*) races (Singh et al. 1991c; Beebe et al. 2000, 2001). All of these races differ in ecological adaptation, geographical range, morpho-agronomic traits, allozyme alleles, and random amplified polymorphic DNA markers (Singh et al. 1991c; Beebe et al. 2000) and their origins are still controversial. It is not known if they are the results of multiple independent domestications within each region, or the result of a single domestication in each region followed by diversification under cultivation.

Indeed, a topic of discussion is whether multiple domestications have occurred within each gene pool and the role of gene flow and introgression. For the Mesoamerican gene pool, different studies have suggested both single (Gepts et al. 1986; Papa and Gepts 2003; Kwak and Gepts 2009; Kwak et al. 2009; Rossi et al. 2009) and multiple domestication events (Singh et al. 1991a, b, c; Beebe et al. 2000; Chacón et al. 2005). In the Andes, the situation is even less clear, because of the lack of geographic structure of the genetic diversity, which reduces the resolving power of the molecular studies. However, both single and multiple domestications have been suggested within the Andean gene pool (Beebe et al. 2001; Santalla et al. 2004; Chacón et al. 2005; Rossi et al. 2009).

Recently, and for the first time, Nanni et al. (2011) approached this question by analyzing nucleotide data, and these strongly support a single domestication event in Mesoamerica. However, the question could not be answered in the Andean gene pool because of the low level of diversity. Using multilocus sequence data to test multiple demographic models in domesticated *P. vulgaris* landraces, Mamidi et al. (2011) suggested that there was a single domestication event in each gene pool. This issue was also undertaken by Bitocchi et al. (2013), by analyzing nucleotide data from five gene fragments, and they clearly indicated a single domestication event for the Mesoamerican gene pool, and suggested a similar scenario for the Andean gene pool.

Other important common bean domestication matters that are still under debate are the identification of the presumed geographic center of domestication, and the domestication dating. Bitocchi et al. (2013) addressed this question and they suggested the Oaxaca valley in Mesoamerica (but see Kwak et al. 2009), and southern Bolivia and northern Argentina in South America, as the origins of common bean domestication.

These results, although encouraging, should be considered with caution, and further efforts are needed to investigate these aspects more deeply, mainly because of the low genetic diversity of the Andean gene pool and because other events might have had roles in shaping the common bean diversity in the areas investigated, such as gene flow between wild and domesticated common bean.

Finally, an important aspect is the occurrence of gene flow between wild and domesticated forms. Using AFLP markers, Papa and Gepts (2003) analyzed the genetic

structure of wild and domesticated populations of *P. vulgaris* from Mexico (with different levels of sympatry). Their results highlighted that the wild and domesticated forms are not genetically isolated, as they show moderate and asymmetric gene flow (> 3-fold higher from domesticated to wild, than *vice versa*). In the presence of gene flow, the marked phenotypic differences between the two forms growing in sympatry are explained by the selection acting against the domesticated alleles in a wild context, and against the wild alleles in an agroecosystem.

Thus, the common bean scenario is characterized by two independent domestication events that gave origins to two clearly differentiated gene pools, and by the co-existence of the wild and domesticated populations, and because crosses between wild and domesticated forms are possible and give fertile and vital progeny, this has made *P. vulgaris* an almost unique and important model among crops for the study of genes and QTLs involved in the domestication process.

20.4 Diffusion and Evolution of *P. vulgaris* Out of the American Centers of Origin

The expansion and the pathways of distribution of the bean out of the American domestication centers were very complex. This also involved several introductions from the New World that were combined with exchanges between continents, and among several countries within continents. In the Old World, the breakdown of the spatial isolation between these two gene pools (Mesoamerican and Andean) increased the potential for their hybridization and introgression. The amplitude of agro-ecological conditions experienced by this crop also dramatically increased, giving new opportunities for both natural and human-mediated selection. Several continents and countries have been proposed as the secondary centers of diversification for *P. vulgaris*, including Europe (Santalla et al. 2002; Angioi et al. 2010, 2011; Gioia et al. [in press](#)), Brazil (Burle et al. 2010), central-eastern and southern Africa (Martin and Adams 1987a, b; Asfaw et al. 2009; Blair et al. 2010) and China (Zhang et al. 2008).

20.4.1 *The Contribution of the Mesoamerican and Andean Gene Pools to Bean Germplasm Collections from Different Parts of the World*

While we must acknowledge that the amount of available data is larger for Europe than for other secondary centers of diversification, it is also clear that the proportions of the Mesoamerican and Andean gene pools can vary considerably across different continents, as also among countries within continents (Fig. 20.2).

Pioneering studies carried out using the phaseolins showed that both the Mesoamerican and Andean gene pools are present in Europe, with a higher frequency

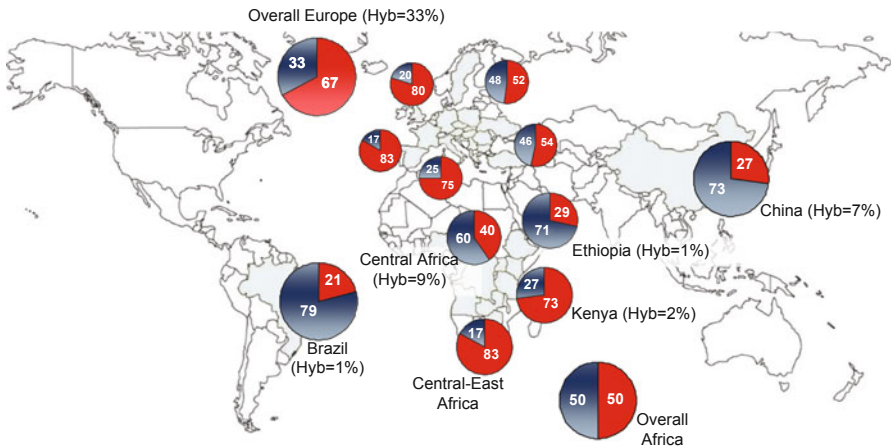


Fig. 20.2 Distribution of the Mesoamerican and Andean gene pools and their hybrids in Europe and other continents. Pie charts show the Mesoamerican (*blue*) and Andean (*red*) gene-pool frequencies (%). Hybrid percentages are indicated next to country names, within brackets. Europe (Angioi et al. 2010) (sample size, 307): Iberian peninsula (53), Italy (32), central-northern Europe (74), eastern Europe (69), south-eastern Europe (79). East Africa (111) (Gepts and Bliss, 1988). Ethiopia (99) and Kenya (89) (Asfaw et al. 2009). Central Africa (355) (Blair et al. 2010). Brazil (279) (Burle et al. 2010). China (299) (Zhang et al. 2008)

for the Andean types (66–76 %; Gepts and Bliss 1988; Lioi 1989), as was subsequently confirmed (76 %) by Logozzo et al. (2007). Recently, to trace the distribution of the domesticated Mesoamerican and Andean gene pools in Europe, Angioi et al. (2010) used cpSSRs, nuclear markers (phaseolin and three indel-spanning markers of *PvSHPI*; Nanni et al. 2011), and morphological seed traits. This study was conducted on a large European collection, and it confirmed that the largest fraction of the European germplasm was of Andean origin (67 %) (Fig. 20.2). The Andean type has been shown to be the most frequent in three European macro-areas: the Iberian peninsula, Italy, and central-northern Europe. The prevalence of the Andean type has also often been confirmed on a local scale (e.g., Limongelli et al. 1996; Escribano et al. 1998; Piergiovanni et al. 2000a, b; Sicard et al. 2005; Angioi et al. 2009b). However, in the eastern part of Europe, the proportion of the Mesoamerican type tends to increase, with a maximum of 46 % in Greece (Fig. 20.2).

Overall, this suggests that there was high gene flow among the different regions of Europe and/or homogeneous selection (either anthropic or ‘natural’). Nonetheless, in some areas, founder effects and/or selection might also have acted. Recently, using methods for the identification of outlier loci for selection, Santalla et al. (2010) provided evidence that selective forces might have had significant roles (particularly for seed size, flowering time, growth habits, pest resistance).

Burle et al. (2010) assessed the genetic diversity and the structure of a sample of 279 geo-referenced common bean landraces from Brazil using nuclear SSR markers, *Phaseolin*, *PvTFLY*, *APA* and *SCAR* markers. They showed that the Mesoamerican and Andean gene pools were both present in Brazil, although the Mesoamerican was four-fold more frequent than the Andean. This is surprising, given the closer

proximity of Brazil to the Andes. To explain these data, Burle et al. (2010) formulated both selection and demographic hypotheses. Similarities in climate and soil between the two areas might explain the success and diffusion of the Mesoamerican bean germplasm in Brazil. Moreover, multiple introductions of Mesoamerican germplasm in pre- and post-conquest times (Gepts et al. 1988) might have had a considerable impact on the establishing of this pattern.

In Africa, the Mesoamerican and Andean gene pools are approximately equal in frequency (Fig. 20.2) (Martin and Adams 1987a; Gepts and Bliss 1988; Asfaw et al. 2009; Blair et al. 2010). However, there are striking differences between different countries. In Kenya (Asfaw et al. 2009), east Africa (Gepts and Bliss 1988) and southern Africa (Martin and Adams 1987a), the Andean type is the most frequent, while in Ethiopia (Asfaw et al. 2009) and central Africa (Blair et al. 2010), the Mesoamerican type predominates. Interestingly, the study of Asfaw et al. (2009) revealed that with some exceptions, the clustering of the accessions was based on the country of origin (Kenya or Ethiopia), with an overall F_{ST} between countries of 0.06 ($P < 0.001$). In particular, the divergence is much greater for the Andean genotypes than for the Mesoamerican ($F_{ST} = 0.34$ and 0.04 , respectively; $P < 0.001$). This suggests that there are at least partially independent seed (and perhaps social) networks in Kenya and Ethiopia, with no strong trans-national bean-seed exchange. An additional reason for the divergence between the Kenyan and Ethiopian germplasm might have arisen through different farmer selection preferences, according to ecological adaptation, cooking value, and market orientation (Wortmann et al. 1998; Asfaw et al. 2009). The predominance of Mesoamerican types in central Africa has been attributed to several reasons: the recent increase in root rot, to which the Andean beans are less resistant (especially determinate types); the higher yield per plant that can often be obtained from Mesoamerica genotypes; and the input of germplasm from national programs (Blair et al. 2010).

China is a large producer of dry beans, and is the most important producer of snap beans in the World, through its intensive horticultural systems that are based on family farms. An analysis of a 229 landraces collection revealed higher prevalence of the Mesoamerican type in China (Zhang et al. 2008). At present, it is believed that there were only a limited number of introductions of the common bean into China (Zheng 1997; Zhang et al. 2008). Thus, one explanation for the prevalence of the Mesoamerican types might be that the few founding populations were biased towards a high frequency of the Mesoamerican type.

20.4.2 The Diversity of the ‘out-of-America’ Germplasm, and Their Divergence from the American Source Population

Under the bottleneck model, it is expected that dissemination from the center of origin will lead to a reduction in genetic diversity. Considering the benchmark as the data obtained by Kwak and Gepts (2009) using nuclear SSR markers to characterize domesticated accessions from the centers of origin, it emerges that overall for the

Table 20.1 Comparison of the SSR nuclear diversity between a native American collection and several samples from around the world

Country	Nuclear SSR marker diversity		Source reference
	Genetic diversity (H)	Loss of diversity (%)	
American reference for domesticated bean	0.63	–	Kwak and Gepts 2009
China	0.54	– 14.3	Zhang et al. 2008
Central Africa	0.62	– 1.6	Blair et al. 2010
Ethiopia + Kenya	0.65	3.2	Asfaw et al. 2009
Ethiopia	0.64	(1.6)	Asfaw et al. 2009
Kenya	0.59	(– 6.3)	Asfaw et al. 2009
Brazil	0.48	– 23.8	Burle et al. 2010
Mean ^a	0.59	– 9.1	

^aThe overall mean has been obtained considering China, Central Africa, Ethiopia + Kenya and Brazil

two gene pools the reduction in diversity has been strong for Brazil, intermediate for China, and low or nearly absent for Africa (Table 20.1). This appears counter intuitive, in that it would be expected that the reduction in diversity is in some way proportional to the distance from the center of origin; i.e., that the reduction in diversity in Brazil would be lower, for example, than in China and Africa. This discrepancy probably arises because the dissemination of *Phaseolus* over the last few centuries was tightly linked to the intense commercial activities and the routes that went all around the world, with the possibility that each continent (and country) has been both source and/or sink of bean germplasm several times, and in different historic periods.

A more comprehensive picture was obtained for Europe by Angioi et al. (2010), through direct comparisons of the levels of diversity between two collections using cpSSR markers: one American and one European. Angioi et al. (2010) concluded that the intensity of the cytoplasmic bottleneck that resulted from the introduction of the common bean into Europe was very low or absent (a loss of cpSSR diversity of ca. 2%).

At the nuclear level, Papa et al. (2006) inferred a much higher loss of diversity consequent to the introduction of the common bean into Europe (ca. 30%). However, Angioi (2006) studied the *PvSHPI* nuclear markers and observed that the number of haplotypes and the genetic diversity were both higher in America than in Europe, supporting the hypothesis of a bottleneck at the nuclear level of greater intensity than for cpSSRs.

The lack of a cytoplasmic bottleneck in Europe is somewhat surprising, because cpSSR markers are very sensitive indicators of such phenomena, due to their uniparental inheritance, hypervariability and haploidy (Provan et al. 2001; Ebert and Peakall 2009; Angioi et al. 2009a). The most likely explanation for this is that the founding common bean populations that colonized Europe were highly variable

in their cytoplasmic DNA or that different releases may have had different source populations.

20.4.3 Introgression Between the Mesoamerican and Andean Gene Pools

Hybrids between the Mesoamerican and Andean gene pools are very important for plant breeding, which often finds the need to recombine Mesoamerican and Andean traits (Johnson and Gepts 1999, 2002). Indeed, hybridization can result in the production of novel genotypes and phenotypes (e.g., seed size, nutritional quality, resistance to pathogens; Angioi et al. 2010; Blair et al. 2010; Santalla et al. 2010) that do not occur in either of the parental taxa. Evolutionary novelty can result either from a combination of different traits from both of the parents, or from traits in a hybrid that transgresses the parental phenotypes (transgressive segregation) (Allendorf and Luikart 2007). Hybridization (with introgression) outside of America had more chances due to the breakdown of the geographical barriers and the isolation that existed between the gene pools in the centers of origin.

In general, distinguishing hybrids at the morphological level is not easy. The use of molecular genetic markers thus greatly simplifies the identification and description of hybrids.

A powerful method for the detection of hybridization events is the integration of cytoplasmic and nuclear analyses (Provan et al. 2001; Ebert and Peakall 2009; Angioi et al. 2009a). Using this approach Angioi et al. (2010) found that at least 33 % of the landraces in the collection were hybrids. Interestingly, in a previous study, and using a different marker system, Santalla et al. (2002) also estimated a high percentage of hybrids in their collection from the Iberian peninsula (25 %). In addition to the molecular results, the individuals identified as hybrids also showed evidence of hybridization from the analysis of seed traits. Indeed, seed size and coat traits tend to vary with the level of introgression between the two gene pools, with relatively good agreement. The complementation of cytoplasmic and nuclear analysis has also been applied with success at local scales in Italy, in the Marche region (12 % hybrids; Sicard et al. 2005) and in Sardinia (4 %; Angioi et al. 2009b).

Moreover, in adopting a maximum likelihood approach, Angioi et al. (2010) estimated that about 11 % of their 'pure' Mesoamerican and Andean individuals (derived from recombination from crosses between parents that belong to the two different gene pools) can be regarded as 'hidden' hybrids. Thus, 44 % of their collection appeared to be derived from at least one hybridization event, with a frequency of hybridization between gene pools ranging from 0.12 to 0.15 % per year.

Several other studies have analyzed hybridization among gene pools (Fig. 20.2) using molecular markers and different statistical approaches. In Brazil, Burle et al. (2010) estimated a hybridization percentage of 4.4 % based on phaseolin analysis, although this was reduced to 0.74 % based on Structure analysis of nuclear SSR markers (Pritchard et al. 2000). In Africa, the identification of hybrids was based on

their intermediate positions between the two gene pools in neighbor-joining trees, and this varied from 1 to 10 % in different countries (Asfaw et al. 2009; Blair et al. 2010). In China, in considering the results of principal coordinate analysis and admixture values based on Structure analysis, Zhang et al. (2008) estimated 7 % as hybrids. As a conclusion, all of these studies based on clustering methods (distance-based, such as neighbor-joining, or model-based, such as Structure) have indicated a number of hybrids as between *ca.* 1 % and 10 %, which is much less than the estimates for Europe.

Such differences in hybrid frequency might be real: the co-occurrence of the two gene pools in the same continent or country does not necessarily imply that they had the potential for hybridization; i.e. the two gene pools might have had different levels of sympatry (so different chances of hybridization) in different places. However, it is also possible that the various molecular approaches have different statistical powers for the detection of hybrids. In particular, using approaches that involve clustering methods, parental type and F1 hybrids can be readily identified if many loci are examined (Allendorf and Luikart 2007). However, to distinguish between F2, backcrosses, or later-generation hybrids with model-based Bayesian methods can be challenging, even if many loci are examined and when divergence between parental populations is high (Vähä and Primmer 2006; Allendorf and Luikart 2007). On the contrary, contrasting cytoplasmic and nuclear markers might lead to the unraveling of not only recent, but also some 'historic' hybridization events between the two gene pools.

20.5 Genomic Tools and Germplasm Collections

Given its phylogenetic position in the Phaseoloids (Stefanovic et al. 2009), the common bean is considered a model organism for comparative legume genomics. It is closely related to other economically important members of the papilionid legumes, including cowpea (*Vigna unguiculata*), pigeon pea (*Vigna radiata*) and soybean (*Glycine max*).

The common bean and soybean diverged nearly 20 million years ago, around the time of the major duplication event in soybean (Lavin et al. 2005; Schlueter et al. 2004). Synteny analysis indicates that most segments of any single common bean linkage group are highly similar to two soybean chromosomes (Galeano et al. 2009). McClean et al. (2010) successfully tested the assumption that the common bean genome is a diploid version of the soybean, with a comparison of all of the mapped genes from bean (McConnell et al. 2010) against all of the scaffold sequences (20 pseudochromosomes) from the soybean genome. For these reasons, *P. vulgaris* has proven to be helpful as a model for understanding the larger soybean genome (about 1,100 Mbp), and a comparative genomics approach to gene discovery is practicable for these two evolutionarily related species.

Due to its importance as a grain legume for the human diet (FAO 2010; Broughton et al. 2003; Carvalho et al. 2012), the value of the common bean is best seen through its role as a societal crop, and its improvement is of constant concern (Singh 2001).

With the aim to create new varieties for farmers and consumers, the international consortium for *Phaseolus* genomics “Phaseomics” was founded, to develop bean genomics, transcriptomics and proteomics (Broughton et al. 2003, <http://www.Phaseolus.net>).

Two common bean whole genome sequences (Mesoamerican and Andean; each about 600 Mbp) will soon be released by a group of US (<http://www.phytozome.net/commonbean.php>) (McClellan et al. 2013) and Ibero-American laboratories (<http://mazorka.langebio.cinvestav.mx/Phaseolus/>). While waiting for this to become available, other methods are needed to develop and facilitate these genomic studies. Here, advanced high-throughput genotyping techniques will provide new insights for association mapping studies in the investigation of variants associated with important traits.

DNA sequences are available for many crops; however, apart from the ongoing model genome projects for *Medicago truncatula* and *Lotus japonicus* (Young et al. 2005), and the recently completed soybean genome (Schmutz et al. 2010), there is comparatively little sequence data for other legumes, including the common bean. Nearly, all of the evidence regarding genetic diversity in the common bean is based on multilocus molecular markers (see Papa et al. 2006, and Acosta-Gallegos et al. 2007, for reviews). Only a few studies have investigated nucleotide diversity in this important crop species, and particularly for wild populations. These have included the sequence diversity of the phaseolin locus in wild accessions (Kami et al. 1995), and of three non-coding regions of the dihydroflavonol 4-reductase and chalcone isomerase genes in landraces and modern cultivars (McClellan et al. 2004; McClellan and Lee 2007). Studies by Nanni et al. (2011) and Bitocchi et al. (2012, 2013) investigated the nucleotide diversity for five different genes from a wide sample of wild and domesticated *P. vulgaris* that is representative of its geographical distribution. A larger amount of sequence data (over 500 genes) was obtained in a study by McConnell et al. (2010) for the two parents of one of the major mapping populations of *P. vulgaris*, ‘BAT93’ × ‘JaloEPP558’ (Freyre et al. 1998). This provided enrichment of the genetic map and allowed investigation of macrosynteny between the common bean and the model organisms of *A. thaliana*, *M. truncatula* and *L. japonicus*. McConnell et al. (2010) exploited over 2,686 *P. vulgaris* contiguous sequences that were generated by Ramirez et al. (2005), from which they obtained useful sequence data for both BAT93 and JaloEPP558 for 534 gene fragments. Of these 534 fragments, 395 were polymorphic between BAT and Jalo, and 300 were mapped and assigned to the 11 linkage groups of *P. vulgaris*. As an important consequence of this study, these markers have become useful for other Mesoamerican × Andean populations.

Linkage maps have been developed from crosses both between and within Mesoamerican and Andean gene pools (see Kelly et al. 2003, for a review). To date, a collection of over 25 linkage maps have been developed in the common bean. Molecular linkage maps are essential for many purposes, such as gene mapping, QTL analysis, linkage disequilibrium analysis, and synteny, and consequently, to

Table 20.2 Relevant publications on the common bean for transcriptome sequencing and bioinformatics analyses

Reference	Source	Method	Accession	Outcome
Ramirez et al. 2005	Nitrogen-fixing root nodules, phosphorus-deficient roots, developing pods, and leaves cDNA libraries	EST sanger sequencing	Negro Jamapa 81, G19833	21,026 ESTs; ca. 8,000 genes
Melotto et al. 2005	19-day-old trifoliolate leaves, 10-day-old shoots, and 13-day-old shoots inoculated with <i>Colletotrichum lindemuthianum</i> cDNA libraries	EST sanger sequencing	SEL 1308	3,126 genes
Tian et al. 2007	Suppression subtractive shoot and root cDNA library in response to phosphorous starvation	EST sanger sequencing	G19833	72 genes
Thibivilliers et al. 2009	Subtractive rust-resistant cDNA library	EST sanger sequencing	Early gallatin	6,202 ESTs
Blair et al. 2011	Drought tolerance and acid-soil tolerance cDNA libraries	EST sanger sequencing	BAT477, G19833	4,219 genes
Kalavacharla et al. 2011	Leaves, flowers, roots and pods cDNA libraries	Next generation sequencing (454 Roche)	BAT93, Sierra	59,295 genes

find genes with particular agronomic and economic traits, for their application to plant breeding.

The availability of large sets of annotated sequences has arisen through the identification, sequencing and validation of gene expression, and these will help in the development of the accurate and complete structural annotation of the common bean genome, and in the identification of the genetic basis of agriculturally important traits. To date, there have been several relevant publications in the common bean regarding transcriptome sequencing and bioinformatics analyses (Table 20.2).

Ramirez et al. (2005) provided an initial platform for the functional genomics of the common bean. They identified almost 8,000 unique genes that were assembled from more than 20,000 ESTs sequenced from various cDNA libraries. These were derived from the Mesoamerican common bean genotype Negro Jamapa 81, and included nitrogen-fixing root nodules, phosphorus-deficient roots, developing pods, and leaves, and from the leaves of the Andean genotype G19833. They showed the utility of mining EST collections in the common bean for SNPs and provided new tools for genomic studies in this species. These sequences have enriched the collection of ESTs for this important crop, and have provided new understanding of bean metabolism, development, and adaptation to stress. The common bean EST

sequences represent the foundation for genome-wide transcript studies, and they are a source of defined molecular markers for mapping bean linkage groups and anchoring physical maps.

Melotto et al. (2005) obtained over 5,000 sequences from three cDNA libraries from a common bean breeding line, from 19-day-old trifoliate leaves, 10-day-old shoots, and 13-day-old shoots inoculated with *Colletotrichum lindemuthianum*. They finally identified 3,126 unigenes, and of these only 314 showed similarity to sequences from the existing database.

Tian et al. (2007) constructed a suppression subtractive cDNA library to identify genes involved in response to phosphorous starvation. They characterized the differentially expressed genes into five functional groups, and by comparison with the GenBank non-redundant database, they were able to further classify 72 genes.

Over 6,000 new common bean ESTs were obtained by Thibivilliers et al. (2009), again using a subtractive cDNA library, which was constructed from a rust-resistant cultivar. As main result, they identified sequences that were up-regulated in response to susceptible and resistant host-pathogen interactions.

Blair et al. (2011) obtained a total of 4,219 unigenes from two cDNA libraries from the drought tolerant Mesoamerican genotype BAT477 and the acid-soil-tolerant Andean genotype G19833.

Several new genomics technologies have emerged in recent years, including next generation sequencing (Mardis 2008), high-throughput marker genotyping, and -omics technologies. These provide powerful tools for the understanding of genome variations in crop species at the DNA, RNA and protein levels, and particularly for nonmodel plant species (Vera et al. 2008). Next-generation sequencing (Mardis 2008) has revolutionized the “-omic era”, allowing the analysis of millions of reads in a very little time and at much reduced cost.

Kalavacharla et al. (2011) provided new genomic information by sequencing a large number of cDNA libraries from different plant tissues using the Roche 454-FLX pyrosequencing platform: leaves, flowers and roots from a common bean cultivar, and pods derived from the BAT93 breeding line, one of the parents of the core common bean mapping populations. They identified 59,295 common bean unigenes, 31,664 of which were newly discovered sequences. In this way they obtained a substantial transcriptome dataset for common bean and increased the number of *P. vulgaris* ESTs deposited in gene bank by 150 %, which is very useful for functional genomics research. They also detected a high number of microsatellites (SSRs): 1,516 and 4,517, in Roche 454-FLX system-derived and genomic sequences, respectively.

All of these efforts have provided significant resources for the discovery of new genes, for the development of molecular markers for future genetic linkage and QTL analyses, and for comparative studies with other legumes. They will also help in the discovery and understanding of the genes that underlie agriculturally important traits in the common bean.

Next-generation sequencing has significantly increased the speed at which SNPs can be discovered. These provide an ideal marker system for genetic research in many crops, which can be used as molecular markers for research. Furthermore, several

high-throughput platforms have been developed that allow rapid and simultaneous genotyping of up to a million SNP markers (Yan et al. 2010).

However, as in other species for which the complete genome sequence is not yet available, in the common bean the use of next-generation sequencing for SNP discovery is much more difficult and costly. On this basis, Hyten et al. (2010) developed a method to improve the number of SNPs in common bean. This system was developed as a multi-tier reduced representation library, and it coupled sequences obtained from the Roche 454 platform (longer reads) with the Illumina genome analyzer (high-throughput) for SNP discovery, for which no whole genome sequence and normalized cDNA libraries are needed. They revealed 3,487 SNPs, 86 % of which were validated with Sanger sequencing.

The study of Cortes et al. (2011) was the first to explore SNP variations for diversity analysis in the common bean. Using KASPar technology (Cuppen 2007), they validated and accessed SNP diversity at 84 gene-based and 10 nongenic loci in a set of 70 genotypes, which included Andean and Mesoamerican accessions previously evaluated for SSRs (Blair et al. 2006b). They found that SNP markers are especially useful for inter-gene-pool comparisons, but not at the intra-gene-pool scale, where SSR markers are efficient (Sicard et al. 2005; Blair et al. 2006b, 2009; Angioi et al. 2009a; Kwak and Gepts 2009). Recently, Blair et al. (2013) developed an Illumina GoldenGate assay for common bean based on conserved legume gene sequences; they tested a total of 768 SNPs, 736 of which gave high quality reads and were scored in a wide sample of *P. vulgaris* accessions. Overall, they found the GoldenGate assay to be a useful genetic tool for rapid analysis of parental combinations, for germplasm studies, and for evaluation of association panels. The genes or genomic regions responsible for traits of interest can be identified either through conventional linkage mapping or through new genetic approaches, such as advanced-backcross QTL analysis, introgression libraries, multi-parent advanced generation intercross populations, and association genetics. These genes can be introgressed or pyramided to develop superior genotypes, using molecular breeding approaches, such as marker-assisted back crossing, marker-assisted recurrent selection, and genome-wide selection.

Genetic resources constitute a rich source of such 'new' genes. Important collections of common bean germplasm are maintained *ex situ* in the gene banks. The online portal Genesys (<http://www.genesys-pgr.org>) supplies information about the accessions of *P. vulgaris* stored in the gene banks. This portal indicates that there are over 83,000 accessions from 138 countries stored in 63 institutions around the World. The main accessions are landraces (ca. 61,000), improved cultivars (> 8,000), breeding materials (> 2,000), and wild forms (> 1,500). The largest and most diverse common bean collection in the world includes over 31,000 accessions, and it is at the *Centro Internacional de Agricultura Tropical* (CIAT) in Colombia. These come from 104 countries, and in particular Mexico, Peru, Colombia and Guatemala, and also from Europe and Africa, and to a lesser extend from Asia. Another large common bean collection is at the United States Department of Agriculture (USDA- ARS) at Washington State University, where over 12,000 accessions from 94 countries are stored. The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in

Gatersleben (Germany) has the largest collection of *Phaseolus* genetic resources in Europe. About 8,000 accessions of *P. vulgaris* are registered, from 69 countries. An important second gene bank for the common bean in Europe is at 'N.I. Vavilov' Research Institute of Plant Industry (VIR, Russia), with 6,000 accessions from 90 countries.

20.6 Conclusions

The data reviewed above show that the improvements in *P. vulgaris* L. are of constant concern both as a societal crop and as a model species for comparative legume genomics. Different studies have been conducted to determine the diversity levels, the origin, the domestication processes and the evolution of this species.

Mesoamerica has been recently proposed to be the origin of *P. vulgaris*. Thus, the wild beans from South America originated through migration from the Mesoamerica populations. Several additional aspects of the evolution and domestication of the common bean have been widely highlighted using genomic tools, including the identification of single domestication events within both gene pools and the characterization of the spread of this crop out of America, with the parallel reduction of the genetic diversity and occurrence of hybridization between gene pools.

Diversity studies based on different molecular markers have highlighted that a large fraction of the bean genome appears to have been under selection during domestication. More data relative to the relationships between the wild and domesticated forms will likewise help in the mining of wild species for novel allelic variations and genes underlying primary agronomic traits.

At present, it appears clear that the forthcoming genome sequence of the common bean, together with the formerly available genomic tools and genetic resources, will become the paradigm to understand the structural and functional diversity of this crop. Alongside, marker-assisted selection programs and high-throughput selection of improved varieties will provide breeders with valuable instruments to achieve effective enhancement of this crop.

References

- Acosta-Gallegos JA, Kelly JD, Gepts P (2007) Pre-breeding in common bean and use of genetic diversity from wild germplasm. *Crop Sci* 47:44–59
- Akibode S, Maredia M (2011) Global and regional trends in production, trade and consumption of food legume crops. SPIA Report department of agricultural, food and resource economics, Michigan State University, East Lansing, MI
- Allendorf FW, Luikart G (2007) Conservation and the genetics of populations. Blackwell, pp 642
- Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. *Nat Rev Genet* 11:697–709
- Angioi SA (2006) Development and use of molecular tools to study the genetic diversity in *Phaseolus vulgaris* L. and *Phaseolus coccineus* L. PhD thesis, University of Turin, Italy

- Angioi SA, Desiderio F, Rau D et al (2009a) Development and use of chloroplast microsatellites in *Phaseolus* spp. and other legumes. *Plant Biol* 11:598–612
- Angioi SA, Rau D, Rodriguez M et al (2009b) Nuclear and chloroplast microsatellite diversity in *Phaseolus vulgaris* L. from Sardinia (Italy). *Mol Breed* 23:413–429
- Angioi SA, Rau D, Attene G et al (2010) Beans in Europe: origin and structure of the European landraces of *Phaseolus vulgaris* L. *Theor Appl Gen* 121:829–843
- Angioi SA, Rau D, Nanni L et al (2011) The genetic make-up of the European landraces of the common bean. *Plant Genet Resour* 9:197
- Anthony JL, Vonder Haar RA, Hall TC (1990) Nucleotide sequence of an alpha-phaseolin gene from *Phaseolus vulgaris*. *Nucleic Acids Res* 18:3396
- Arumuganthan K, Earle E (1991) Nuclear DNA content of some important plant species. *Plant Mol Biol Rep* 9:208–218
- Asfaw A, Blair MW, Almekinders C (2009) Genetic diversity and population structure of common bean (*Phaseolus vulgaris* L.) landraces from the East African highlands. *Theor Appl Gen* 120:1–12
- Beaver JS, Osorno JM (2009) Achievements and limitations of contemporary common bean breeding using conventional and molecular approaches. *Euphytica* 168:145–175
- Becerra-Velásquez VL, Gepts P (1994) RFLP diversity in common bean (*Phaseolus vulgaris* L.). *Genome* 37:256–263
- Beebe S, Skroch P, Tohne J et al (2000) Structure of genetic diversity among common bean landraces of middle-American origin based on correspondence analysis of RAPD. *Crop Sci* 40:264–273
- Beebe S, Rengifo J, Gaitan E et al (2001) Diversity and origin of Andean landraces of common bean. *Crop Sci* 41:854–862
- Bennett MD, Leitch IJ (1995) Nuclear DNA amounts in angiosperms. *Ann Bot* 76:113–116
- Bennett MD, Leitch IJ (2010) Angiosperm DNA C-values database (release 7.0, Dec. 2010) <http://www.kew.org/cvalues/>
- Bitocchi E, Nanni L, Bellucci E et al (2012) Mesoamerican origin of the common bean (*Phaseolus vulgaris* L.) is revealed by sequence data. *Proc Natl Acad Sci U S A* 109(14):E788–E796
- Bitocchi E, Bellucci E, Giardini A et al (2013) Molecular analysis of the parallel domestication of the common bean in Mesoamerica and the Andes. *New Phytol* 197:300–313
- Blair MW, Iriarte G, Beebe S (2006a) QTL analysis of yield traits in an advanced backcross population derived from a cultivated Andean x wild common bean (*Phaseolus vulgaris* L.) cross. *Theor Appl Gen* 112:1149–1163
- Blair MW, Giraldo MC, Buendia HF et al (2006b) Microsatellite marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor Appl Gen* 113:100–109
- Blair MW, Diaz LM, Buendia HF et al (2009) Genetic diversity, seed size associations and population structure of a core collection of common beans (*Phaseolus vulgaris* L.). *Theor Appl Genet* 119:955–972
- Blair MW, González LF, Kimani M et al (2010) Genetic diversity, inter-gene pool introgression and nutritional quality of common beans (*Phaseolus vulgaris* L.) from Central Africa. *Theor Appl Gen* 121:237–248
- Blair MW, Fernandez AC, Ishitani M et al (2011) Construction and EST sequencing of full-length, drought stress cDNA libraries for common beans (*Phaseolus vulgaris* L.). *BMC Plant Biol* 11:171
- Blair MW, Soler A, Cortés AJ (2012) Diversification and population structure in common beans (*Phaseolus vulgaris* L.). *PLoS One* 7(11):e49488
- Blair MW, Cortés AJ, Penmetza RV et al (2013) A high-throughput SNP marker system for parental polymorphism screening, and diversity analysis in common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 126:535–548
- Broughton WJ, Hernandez G, Blair M et al (2003) Beans (*Phaseolus* spp.)—model food legumes. *Plant Soil* 252:55–128
- Burle ML, Fonseca JR, Kami JA et al (2010) Microsatellite diversity and genetic structure among common bean (*Phaseolus vulgaris* L.) landraces in Brazil, a secondary center of diversity. *Theor Appl Genet* 121:801–813
- Carvalho LMJ, Correa MM, Pereira EJ et al (2012) Iron and zinc retention in common beans (*Phaseolus vulgaris* L.) after home cooking. *Food Nut Res* 56:15618

- Chacón SMI, Pickersgill B, Debouck DG (2005) Domestication patterns in common bean (*Phaseolus vulgaris* L.) and the origin of the Mesoamerican and Andean cultivated races. *Theor Appl Genet* 110:432–444
- Coates AG, Collins LS, Aubry MP et al (2004) The geology of the Darien, Panama, and the late Miocene–Pliocene collision of the Panama arc with north–western South America. *Geol Soc Amer Bull* 116:1327–1344
- Cortés AJ, Chavarro MC, Blair MW (2011) SNP marker diversity in common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 123:827–845
- Cuppen E (2007) Genotyping by allele–specific amplification (KASPar). *Cold Spring Harb Protocols*, pp 172–173
- Davey JW, Hohenlohe PA, Etter PD et al (2011) Genome–wide genetic marker discovery and genotyping using next–generation sequencing. *Nat Rev Genet* 12:499–510
- Debouck DG, Toro O, Paredes OM et al (1993) Genetic diversity and ecological distribution of *Phaseolus vulgaris* in northwestern South America. *Econ Bot* 47:408–423
- Delgado-Salinas A, Turley T, Richman A et al (1999) Phylogenetic analysis of the cultivated and wild species of *Phaseolus* (Fabaceae). *Syst Bot* 24:438–460
- Delgado-Salinas A, Bibler R, Lavin M (2006) Phylogeny of the genus *Phaseolus* (Leguminosae): a recent diversification in an ancient landscape. *Syst Bot* 31:779–791
- Desiderio F, Bitocchi E, Bellucci E et al (2013) Chloroplast microsatellite diversity in *Phaseolus vulgaris*. *Front Plant Sci* 3:312
- Díaz LM, Blair MW (2006) Race structure within the Mesoamerican gene pool of common bean (*Phaseolus vulgaris* L.) as determined by microsatellite markers. *Theor Appl Genet* 114:143–154
- Ebert D, Peakall R (2009) Chloroplast simple sequence repeats (cpSSRs): technical resources and recommendations for expanding cpSSR discovery and applications to a wide array of plant species. *Mol Ecol Resour* 9:673–690
- Escribano MR, Santalla M, Casquero PA et al (1998) Patterns of genetic diversity in landraces of common bean (*Phaseolus vulgaris* L.) from Galicia. *Plant Breed* 117:49–56
- Freyre R, Ríos R, Guzmán L et al (1996) Ecogeographic distribution of *Phaseolus* spp. (Fabaceae) in Bolivia. *Econ Bot* 50:195–215
- Freyre R, Skroch P, Geffroy V et al (1998) Towards an integrated linkage map of common bean. 4. Development of a core map and alignment of RFLP maps. *Theor Appl Genet* 97:847–856
- Freytag GF, Debouck DG (1996) *Phaseolus costaricensis*, a new wild bean species (Phaseolinae, Leguminosae) from Costa Rica and Panama, central America. *Novon* 6:157–163
- Freytag GF, Debouck DG (2002) Taxonomy, distribution, and ecology of the genus *Phaseolus* (Leguminosae–Papilionoideae) in North America, Mexico and central America. Botanical Research Institute of Texas, Ft. Worth
- Galeano CH, Fernandez AC, Gomez M et al (2009) Single strand conformation polymorphism based SNP and indel markers for genetic mapping and synteny analysis of common bean (*Phaseolus vulgaris* L.). *BMC Genomics* 10:629
- Galeano C, Cortés A, Fernández A et al (2012) Gene–based single nucleotide polymorphism markers for genetic and association mapping in common bean. *BMC Genet* 13(1):48
- Gepts P, Bliss FA (1985) F1 hybrid weakness in the common bean: differential geographic origin suggests two gene pools in cultivated bean germplasm. *J Hered* 76:447–450
- Gepts P, Osborn TC, Rashka K et al (1986) Phaseolin–protein variability in wild forms and landraces of the common bean (*Phaseolus vulgaris*): evidence for multiple centers of domestication. *Econ Bot* 40:451–468
- Gepts P, Bliss FA (1988) Dissemination pathways of common bean (*Phaseolus vulgaris*, Fabaceae) deduced from phaseolin electrophoretic variability. II Europe and Africa. *Econ Bot* 42:86–104
- Gepts P, Kmiecik K, Pereira P et al (1988) Dissemination pathways of common bean (*Phaseolus vulgaris*, Fabaceae) deduced from phaseolin electrophoretic variability. I. The Americas. *Econ Bot* 42:73–85
- Gepts P, Debouck DG (1991) Origin, domestication, and evolution of the common bean, *Phaseolus vulgaris*. In: Voyses O, Van Schoonhoven A (eds.) *Common beans: research for crop improvement*. CAB, Oxon, UK, pp 7–53
- Gepts P (1998) Origin and evolution of common bean, past event and recent trends. *J Am Soc Hortic Sci* 33:1124–1130

- Gepts P, Papa R, Coulibaly S et al (1999) Wild legume diversity and domestication – insights from molecular methods. In Vaughan D (ed), Wild legumes, Proc. 7th MAFF International Workshop on Genetic Resources. National Institute of Agrobiological Resources, Tsukuba, Japan, pp 19–31
- Gepts P, Papa R (2002). Evolution during domestication. In: Encyclopedia of Life Sciences 1–7 LONDON: Nature Publishing Group. Macmillan Publishers Ltd (UK)
- Gepts P, Aragão F, de Barros E et al (2008) Genomics of *Phaseolus* beans, a major source of dietary protein and micronutrients in the Tropics. In: Moore PH, Ming R (eds) Genomics of Tropical Crop Plants. Springer, Berlin, pp 113–143
- Gioia T, Logozzo G, Attene G et al (2013) Evidence for introduction bottleneck and extensive inter-gene pool (Mesoamerica x Andes) hybridization in the European common bean (*Phaseolus vulgaris* L.) germplasm Plos ONE (in press)
- Glémin S, Bataillon T (2009) A comparative view of the evolution of grasses under domestication. New Phytol 183:273–290
- Gupta PK, Rustgi S, Mir RR (2008) Array-based high-throughput DNA markers for crop improvement. Heredity 101:5–18
- Hyten DL, Song Q, Fickus EW et al (2010) High-throughput SNP discovery and assay development in common bean. BMC Genomics 11:475
- Johnson WC, Gepts P (1999) Segregation for performance in recombinant inbred populations resulting from inter-gene pool crosses of common bean (*Phaseolus vulgaris* L.). Euphytica 106:5–56
- Johnson WC, Gepts P (2002) The role of epistasis in controlling seed yield and other agronomic traits in an Andean–Mesoamerican cross of common bean (*Phaseolus vulgaris* L.). Euphytica 125:69–79
- Kalavacharla V, Liu Z, Meyers BC et al (2011) Identification and analysis of common bean (*Phaseolus vulgaris* L.) transcriptomes by massively parallel pyrosequencing. BMC Plant Biol 11:135
- Kami J, Becerra–Velásquez V, Debouck DG et al (1995) Identification of presumed ancestral DNA sequences of phaseolin in *Phaseolus vulgaris*. Proc Natl Acad Sci U S A 92:1101–1104
- Kelly JD, Gepts P, Miklas PN et al (2003) Tagging and mapping of genes and QTL and molecular-marker assisted selection for traits of economic importance in bean and cowpea. Field Crops Res 82:135–154
- Koenig R, Gepts P (1989) Allozyme diversity in wild *Phaseolus vulgaris*: further evidence for two major centers of diversity. Theor Appl Genet 78:809–817
- Koinange EMK, Singh SP, Gepts P (1996) Genetic control of the domestication syndrome in common bean. Crop Sci 36:1037–1145
- Kwak M, Kami JA, Gepts P (2006) Identification of the determinacy gene (Fin) and its evolution during domestication in common bean (*Phaseolus vulgaris* L.). In: Plant & Animal Genome XIV, poster 447. Abstract available at: http://www.intl-pag.org/14/abstracts/PAG14_P447.html
- Kwak M, Velasco D, Gepts P (2008) Mapping homologous sequences for determinacy and photoperiod sensitivity in common bean (*Phaseolus vulgaris*). J Hered 99:283–291
- Kwak M, Gepts P (2009) Structure of genetic diversity in the two major gene pools of common bean (*Phaseolus vulgaris* L., Fabaceae). Theor Appl Genet 118:979–992
- Kwak M, Kami JA, Gepts P (2009) The putative Mesoamerican domestication center of *Phaseolus vulgaris* located in the Lerma–Santiago basin of Mexico. Crop Sci 49:554–563
- Lavin M, Herendeen PS, Wojcickowski MF (2005) Evolutionary rate analysis of leguminosae implicates a rapid diversification of lineages during the tertiary. Syst Biol 54:575–594
- Limongelli G, Laghetti G, Perrino P et al (1996) Variation of seed storage protein in landraces of common bean (*Phaseolus vulgaris* L.) from Basilicata, southern Italy. Plant Breed 119:513–516
- Lioi L (1989) Geographical variation of phaseolin patterns in an old world collection of *Phaseolus vulgaris*. Seed Sci Technol 17:317–324
- Logozzo G, Donnoli R, Macaluso L et al (2007) Analysis of the contribution of Mesoamerican and Andean gene pools to European common bean (*Phaseolus vulgaris* L.) germplasm and strategies to establish a core collection. Genet Resour Crop Ev 54:1763–1779
- Mamidi S, Rossi M, Annam D et al (2011) Investigation of the domestication of common bean (*Phaseolus vulgaris*) using multilocus sequence data. Funct Plant Biol 38:953–967

- Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet* 24:133–141
- Marshall HD, Newton C, Ritland K (2002) Chloroplast phylogeography and evolution of highly polymorphic microsatellites in lodgepolepine (*Pinus contorta*). *Theor Appl Genet* 104:367–378
- Martin GB, Adams MW (1987a) Landraces of *Phaseolus vulgaris* (Fabaceae) in northern Malawi I. Regional variation. *Econ Bot* 41:190–203
- Martin GB, Adams MW (1987b) Landraces of *Phaseolus vulgaris* (Fabaceae) in northern Malawi II. Generation and maintenance of variability. *Econ Bot* 41:204–215
- McClellan PE, Lee RK, Miklas PN (2004) Sequence diversity analysis of dihydroflavonol 4-reductase intron 1 in common bean. *Genome* 47:266–280
- McClellan PE, Lee RK (2007) Genetic architecture of chalcone isomerase non-coding regions in common bean (*Phaseolus vulgaris* L.). *Genome* 50:203–214
- McClellan PE, Lavin M, Gepts P et al (2008) *Phaseolus vulgaris*: a diploid model for soybean. In: Stacey G (eds) *Soybean Genomics*. Springer, Berlin, pp 55–78
- McClellan PE, Mamidi S, McConnell M et al (2010) Syntenic mapping between common bean and soybean reveals extensive blocks of shared loci. *BMC Genomics* 11:184. <http://www.biomedcentral.com/1471-2164/11/184>
- McClellan PE, Jackson S, Schmutz J et al (2013) Progress toward a draft sequence of the common bean genome. *Grains and Legumes* (in press)
- McConnell M, Mamidi S, Lee R et al (2010) Syntenic relationships among legumes revealed using a gene-based genetic linkage map of common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 121:1103–1116
- McCouch S (2004) Diversifying selection in plant breeding. *PLoS Biol* 2:e347
- Melotto M, Monteiro-Vitorello CB, Bruschi AG et al (2005) Comparative bioinformatic analysis of genes expressed in common bean (*Phaseolus vulgaris* L.) seedlings. *Genome* 48:562–570
- Miklas PN, Kelly JD, Beebe SE et al (2006) Common bean breeding for resistance against biotic and abiotic stresses: from classical to MAS breeding. *Euphytica* 147:105–131
- Nanni L, Bitocchi E, Bellucci E et al (2011) Nucleotide diversity of a genomic sequence similar to SHATTERPROOF (PvSHP1) in domesticated and wild common bean (*Phaseolus vulgaris* L.). *Theor Appl Genet* 123:1341–1357
- Papa R, Gepts P (2003) Asymmetry of gene flow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. *Theor Appl Genet* 106:239–250
- Papa R, Acosta J, Delgado-Salinas A et al (2005) A genome-wide analysis of differentiation between wild and domesticated *Phaseolus vulgaris* from Mesoamerica. *Theor Appl Genet* 111:1147–1158
- Papa R, Nanni L, Sicard D et al (2006) The evolution of genetic diversity in *Phaseolus vulgaris* L. In: Motley TJ, Zerega N, Cross H (eds) *New Approaches to the Origins, Evolution and Conservation of Crops: Darwin's Harvest*. Columbia University Press, New York
- Papa R, Bellucci E, Rossi M et al (2007) Tagging the signatures of domestication in common bean (*Phaseolus vulgaris*) by means of pooled DNA samples. *Ann Bot* 100:1039–1051
- Piergiovanni AR, Cerbino D, Brandi M (2000a) The common bean populations from Basilicata (southern Italy). An evaluation of their variation. *Genet Resour Crop Ev* 47:489–495
- Piergiovanni AR, Taranto G, Pignone D (2000b) Diversity among common bean populations from the Abruzzo region (central Italy): a preliminary inquiry. *Genet Resour Crop Ev* 47:467–470
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Provan J, Soranzo N, Wilson NJ et al (1999) A low mutation rate for chloroplast microsatellites. *Genetics* 153:943–947
- Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol Evol* 16:142–147
- Ramirez M, Graham MA, Blanco-Lopez L et al (2005) Sequencing and analysis of common bean ESTs. Building a foundation for functional genomics. *Plant Physiol* 137:1211–1227
- Repinski S, Kwak M, Gepts P (2012) The common bean growth habit gene PvTFL1y is a functional homolog of Arabidopsis TFL1. *Theor Appl Genet* 124:1539–1547
- Rossi M, Bitocchi E, Bellucci E et al (2009) Linkage disequilibrium and population structure in wild and domesticated populations of *Phaseolus vulgaris* L. *Evol Appl* 2:504–522

- Santalla M, Rodiño AP, De Ron AM (2002) Allozyme evidence supporting southwest Europe as a secondary center of genetic diversity for common bean. *Theor Appl Genet* 104:934–944
- Santalla M, Menéndez-Sevillano MC, Monteagudo AB et al (2004) Genetic diversity of Argentinean common bean and its evolution during domestication. *Euphytica* 135:75–87
- Santalla M, De Ron AM, De La Fuente M (2010) Integration of genome and phenotypic scanning gives evidence of genetic structure in Mesoamerican common bean (*Phaseolus vulgaris* L.) landraces from the southwest of Europe. *Theor Appl Genet* 120:1635–1651
- Schlueter JA, Dixon P, Granger C et al (2004) Mining the EST databases to determine evolutionary events in the legumes and grasses. *Genome* 47:868–876
- Schmutz J, Cannon SB, Schlueter J et al (2010) Genome sequence of the paleopolyploid soybean. *Nature* 463:178–183
- Sicard D, Nanni L, Porfiri O et al (2005) Genetic diversity of *Phaseolus vulgaris* L and *P. coccineus* L. landraces in central Italy. *Plant Breed* 124:464–472
- Singh SP (2001) Broadening the genetic base of common bean cultivars. *Crop sci* 41:1659–1675
- Singh SP, Nodari R, Gepts P (1991a) Genetic diversity in cultivated common bean. I. Allozymes. *Crop Sci* 31:19–23
- Singh SP, Gutiérrez JA, Molina A et al (1991b) Genetic diversity in cultivated common bean. II. Marker-based analysis of morphological and agronomic traits. *Crop Sci* 31:23–29
- Singh SP, Gepts P, Debouck DG (1991c) Races of common bean (*Phaseolus vulgaris* L., Fabaceae). *Econ Bot* 45:379–396
- Stefanović S, Pfeil BE, Palmer JD et al (2009) Relationships among phaseoloid legumes based on sequences from eight chloroplast regions. *Syst Bot* 34:115–128
- Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. *Science* 277:1063–1066
- Thibivilliers S, Joshi T, Campbell KB et al (2009) Generation of phaseolus vulgaris ESTs and investigation of their regulation upon uromyces appendiculatus infection. *BMC Plant Biol* 9:46
- Thuillet AC, Bataillon T, Poirier S et al (2005) Estimation of long-term effective population sizes through the history of durum wheat using microsatellite data. *Genetics* 169:1589–1599
- Tian J, Venkatachalam P, Liao H et al (2007) Molecular cloning and characterization of phosphorous starvation responsive genes in common bean (*Phaseolus vulgaris* L.). *Planta* 227:151–165
- Tohme J, Gonzalez DO, Beebe S et al (1996) AFLP analysis of gene pools of a wild bean core collection. *Crop Sci* 36:1375–1384
- Toro O, Tohme J, Debouck DG (1990) Wild bean (*Phaseolus vulgaris* L.): Description and distribution. Centro Internacional de Agricultura Tropical, Cali, Colombia
- Tuberosa R, Graner A, Varshney RK (2011) Genomics of plant genetic resources: an introduction. *Plant Genet Resour* 9:151–154
- Vähä JP, Primmer CR (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Mol Ecol* 15:63–72
- Heerwaarden J van, Doebley J, Briggs WH et al (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc Natl Acad Sci U S A* 108:1088–1092
- Vera J, Wheat C, Fescemyer H et al (2008) Rapid transcriptome characterization for a non model organism using 454 pyrosequencing. *Mol Ecol* 17:1636–1647
- Wortmann CS, Kirkby RA, Eledu CA et al (1998) Atlas of common bean (*Phaseolus vulgaris* L.) production in Africa. CIAT Pan-African Bean Research Alliance, vol 133
- Yan JB, Yang XH, Shah T et al (2010) High-throughput SNP genotyping with the GoldenGate assay in maize. *Mol Breed* 25:441–451
- Young ND, Cannon SB, Sato S et al (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Phys* 137:1174–1181
- Zhang X, Blair MW, Wang S (2008) Genetic diversity of Chinese common bean (*Phaseolus vulgaris* L.) landraces assessed with simple sequence repeats markers. *Theor Appl Genet* 117:629–640
- Zheng ZJ (1997) Food legumes in China. China Agriculture Press, Beijing, pp 222–249

Part IV
Mining Allelic Diversity

Chapter 21

Advances in *Nicotiana* Genetic and “Omics” Resources

James N. D. Battey, Nicolas Sierro, Nicolas Bakaher and Nikolai V. Ivanov

Contents

21.1	Introduction	512
21.1.1	Species, Phylogeny and Evolution of the Genus <i>Nicotiana</i>	512
21.1.2	Relevance of the Genus <i>Nicotiana</i>	513
21.2	Genetics	515
21.3	Genome	516
21.3.1	<i>N. benthamiana</i> Genome Draft	517
21.3.2	Tobacco Genome Initiative (<i>N. tabacum</i> and <i>N. benthamiana</i>)	518
21.3.3	PMI Physical Map Effort (<i>N. tabacum</i>)	518
21.3.4	PMI Whole Genome Shotgun Effort (<i>N. tabacum</i>)	519
21.3.5	Transcription Control	520
21.3.6	Epigenomics	521
21.4	Transcriptome	521
21.4.1	Expressed Sequence Tag Libraries	522
21.4.2	Microarray Analyses	522
21.4.3	RNA-Seq Effort	524
21.4.4	Small RNA Data	524
21.5	Proteome	525
21.5.1	Protein Resources and Prediction From the Genome	525
21.5.2	Proteomics Characterization Using Tandem Mass Spectrometry	525
21.6	Metabolome	526
21.6.1	Metabolic Network Resources	526
21.6.2	Application of Metabolomics Information	527
21.7	Summary	527
	References	528

Abstract The importance of the genus *Nicotiana* is both economic, as it contains species which are major cash crops, and scientific, as many of the species’ genomes are highly complex due to their evolutionary history. Investigating these species has been accelerated by advances in “Omics” techniques; these high-throughput methods, nucleic acid sequencing in particular, have led to an explosion in the volume of data available.

N. V. Ivanov (✉) · J. N. D. Battey · N. Sierro · N. Bakaher
Philip Morris International R&D, Philip Morris Products SA, Neuchâtel, Switzerland
email: nikolai.ivanov@pmi.com

Here we provide an overview of how these data are organized and stored as public genome resources, as well as how *Nicotiana* researchers can mine or otherwise use these data to answer scientific questions, such as by constructing microarrays from expressed sequence tag (EST) and genomic data. We further examine past sequencing efforts, such as the Tobacco Genome Initiative (TGI), and the objectives and progress of current projects in the field, in particular for *N. tabacum*, *N. sylvestris*, *N. tomentosiformis*, and *N. benthamiana*, which are part of the SOL100 target list. We look at how these data can be leveraged in the future, as for example by using genomic sequences for proteogenomics, or creating species-specific metabolic pathway repositories generated from genome annotation data. *Nicotiana* genetic and “Omics” resources can be used to improve breeding strategies to obtain desirable traits such as disease and stress resistance, yield and quality.

Keywords *Nicotiana* · Tobacco · Bioinformatics · Genetics · Genomics · Transcriptomics · Proteomics · Metabolomics · Genetic marker · Genetic map

21.1 Introduction

21.1.1 *Species, Phylogeny and Evolution of the Genus Nicotiana*

The Solanaceae family has its evolutionary and geographical origin in South America, though it has spread from there to Africa and beyond (Eich 2008). The family is of great agronomic importance as it contains many of the world’s staple crops, potatoes, tomatoes and peppers being but three examples. “Omics” studies promise to accelerate research and plant breeding, and have therefore become a priority amongst researchers. For example, the SOL100 initiative represents a coordinated effort by the community to obtain the genome sequence of many Solanaceae species. The group of *Nicotiana* is a genus of the Solanaceae plant family, which contains another economically important species, tobacco. Recent genomic sequencing efforts have shown that some *Nicotiana* genomes are large, at least compared to other Solanaceae such as the tomato (The Tomato Genome Consortium 2012).

Three subgenera of the genus *Nicotiana* have been described by Goodspeed (Goodspeed 1954) based on morphological description, chromosome number and geographical distribution: *rustica* (e.g. *N. rustica*), *tabacum* (e.g., *N. tomentosiformis*, *N. tabacum*) and *petunioides* (e.g., *N. sylvestris*, *N. glauca*). These subgenera have been confirmed using genomic *in situ* hybridization (GISH) and molecular genetics (Chase et al. 2003; Khan and Narayan 2007) and the classification has been slightly improved in some of the subsections and now includes *Nicotiana* spp. discovered after Goodspeed’s publication, like *N. kawakamii* discovered in 1968 in Bolivia and described by Ohashi in 1976 or *N. africana* described by Merxmüller and Buttler in 1975.

Several members of the *Nicotiana* genus are particularly interesting from an evolutionary point of view. As the result of relatively recent allotetraploidization events, they effectively harbor two genomes, which they have inherited from their diploid ancestors. Using such species, one can study the processes which follow such cases of reticulate evolution (Kelly et al. 2010), such as genome reduction (Leitch et al. 2008) and the alteration of expression patterns to compensate for the gene redundancy incurred. For example, *N. benthamiana* is thought to have originated via allotetraploidization of ancestors from the sections *Sylvestres* and *Noctiflorae*. *N. tabacum* is presumed to be the result of the hybridization of *N. sylvestris* as the maternal donor and *N. tomentosiformis* as the paternal donor (Ren and Timko 2001), but it has also been proposed that the paternal donor may have been an introgression hybrid of *N. tomentosiformis* and *N. otophora* (Murad et al. 2002).

N. tabacum is particularly well characterized due to its commercial importance. The U.S. *Nicotiana* Germplasm Collection (Nicholson et al. 2009) contains many *N. tabacum* varieties with around 1900 accessions. Many other genetic resources have been developed by farmers and by breeders all over the globe. Although in some countries like the U.S. there is a history of tobacco breeding starting already in the 19th century, in many other countries imported tobacco seeds were kept and crossed by farmers over the years. These farmer seeds have spread locally, and specific genetic patterns can be identified while screening this material with sufficient genetic markers, as has been shown for Macedonian tobacco (Davalieva et al. 2010), Thai tobacco (Denduangboripant et al. 2010), Chinese flue-cured tobacco (Liu et al. 2009) and Indian tobacco (Sarala and Rao 2008).

There are eight market classes of commercial tobacco: burley, cigar filler, cigar wrapper, dark air-cured, dark fire-cured, flue-cured, Maryland, and oriental (Nicholson et al. 2009). The tobacco accessions which are not cultivated for industry purpose are often referred to as “primitives”. They represent the majority of the genetic diversity of *N. tabacum*. Modern breeding of tobacco has created a severe bottleneck in the allelic diversity of tobacco (Moon et al. 2004), mostly for the flue-cured and burley market types. The recent profiling of 312 tobacco accessions with 49 SSR markers showed that the global population of tobacco is highly structured (Fricano et al. 2012) and this structure corresponds well to the market classes present in this subsample. A subset the U.S. *Nicotiana* Germplasm Collection was genotyped using simple sequence repeat (SSR) markers. Although it is easy to identify the main market classes (burley and flue-cured), there are many conflicts between the classifications according to the database and the classifications according to the measured genetic distance (see Fig. 21.1).

21.1.2 Relevance of the Genus *Nicotiana*

The genus *Nicotiana* is of scientific interest for several reasons. Firstly, its economic importance has made it a target for selective breeding for the purpose of crop improvement (e.g., for yield, or resistance to stress or pathogens). Secondly, many

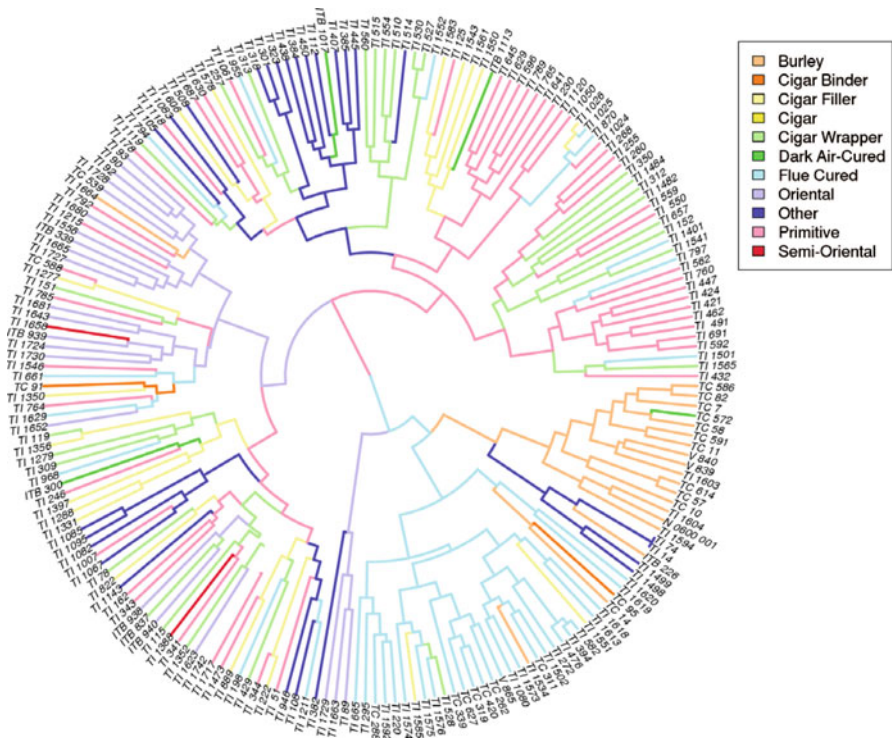


Fig. 21.1 Phylogenetic tree based on genotyping data of 206 selected tobacco accessions from the U.S. *Nicotiana* Germplasm Collection. The Burley and Flue-cured varieties form clear and closely related subclusters; other types are not as clearly defined

Nicotiana species are known to have high levels of secondary metabolite production. They produce a range of alkaloids such as anatabine, anabasine, nornicotine and nicotine, which are thought to act as a deterrent to herbivores. Potentially, studying the secondary metabolites of *Nicotiana* species could identify new bioactive compounds with medicinal properties. Thirdly, their use as model organisms gives insights into plant biology, which are useful far beyond the genus, even beyond the family itself. Tobacco's short generation times of 87-138 days (Lewis and Kernodle 2009) and high seed yield have made it a useful tool for plant biologists. *N. benthamiana*, another noteworthy example, is a model organism for plant infection biology, due to its susceptibility to many diseases (Goodin et al. 2008). Additionally, many *Nicotiana* species are readily amenable to transformation, making them very accessible for genetic manipulation and experimentation. This ease of manipulation also makes *Nicotiana* species attractive as tools for expressing therapeutic molecules, such as the antimalarial compound artemisin (Farhi et al. 2011) and they have been identified as potential vaccine production platforms (Ling et al. 2012). Moreover, the BY-2 cell line from tobacco (Nagata et al. 2004) is one of the most widely used cell lines in experimental plant research. This cell line has proven to be particularly useful for

understanding the plant cell cycle, because cell division can be readily synchronized in BY-2 cultures (Geelen and Inze 2001). Finally, as key plants in the genus are allotetraploids, they can be used to study the processes involved in reticulate evolution: tobacco for example has undergone genome downsizing events after polyploidization (Renny-Byfield et al. 2011) and a detailed genomic and transcriptomic analysis would shed light on the exact gene deletion and regulation changes that compensate for gene duplication.

21.2 Genetics

A genetic map contains information on the relative location of genetic markers based on their recombination frequency. The more frequently two elements recombine, the more distant they are physically on a chromosome. These maps can therefore be used to study the physical evolution of chromosomes or perform Quantitative Trait Loci (QTL) experiments. The markers are used for genetic diversity studies as described above or for Marker-Assisted selection in order to accelerate breeding or to link sequencing and physical map data to pseudo-chromosomes (linkage group).

Genetic maps of *N. tomentosiformis* (*N. tomentosiformis* TA3385 x *N. otophora* TA3353) and *N. acuminata* (*N. acuminata* TA3460 x *N. acum multiflora* TA3461), a species close to *N. sylvestris*, have been constructed using single-copy conserved ortholog (COSII) markers developed for euasterid species (Wu et al. 2006) and SSR markers from tobacco (Bindler et al. 2007). The obtained *N. tomentosiformis* genetic map contains 262 COSII and 221 SSR markers, and the *N. acuminata* genetic map 133 COSII and 174 SSR markers (Wu et al. 2010). Combining these maps with those from tomato and tobacco indicates where rearrangements have occurred during evolution, which have been found to be more frequent in tobacco than in its relatives. These two maps are also useful tools to study the syntenic relationship between the Nicotiana genus and other Solanaceae such as tomato, eggplant and pepper.

Extending the previous marker-based linkage map (Bindler et al. 2007), a high-density genetic map of *N. tabacum* was generated using an F2 mapping population derived from the intervarietal cross of Hicks Broadleaf and Red Russian. This map contains 2,317 SSR markers and 2,363 loci (Bindler et al. 2011). Amplification of SSR markers in *N. tomentosiformis* and *N. sylvestris* was used to determine the possible origin of chromosomes or parts thereof. Around 50% of the markers amplified both S and T genome showing a relatively low level of divergence since the hybridization. The 2007 version, based on the data from (Bindler et al. 2007), of the genetic map for tobacco is available on the SOL genomics website (http://solgenomics.net/cview/map.pl?map_id=15) and the 2011 version, based on the data in reference (Bindler et al. 2011), is anticipated to be added soon.

The first example of a QTL mapping approach in tobacco focused on leaf and smoke properties and was published by Julio et al. (Julio et al. 2006). The low level of polymorphism in the population and the marker types used (AFLP; ISSR, SCAR and SSAP) led to a partial genetic map and substantial segregation distortion.

Nevertheless, seventy-five QTLs associated to physical and chemical properties of tobacco leaves and smoke were identified on twelve linkage groups. The availability of a dense SSR genetic map facilitates quantitative genetic studies in tobacco and improves tracking of introgressions. Microsatellite markers were used in QTL mapping experiments (Vontimitta et al. 2010; Vontimitta and Lewis 2012) in order to map genes related to leaf surface components (cis-abienol and sucrose esters) that likely influence plant resistance to pests, or to map QTLs linked to black shank (*Phytophthora nicotianae*) resistance. The major goal of this research is to identify QTLs and associated genetic markers for tobacco crop improvement towards disease or pest resistance. So far, in tobacco, many of the disease resistance were introgressed from *Nicotiana spp.* They often consists of single locus: for example blue mold resistance from a *Nicotiana* of Australian origin (Milla et al. 2005), Tobacco mosaic virus (TMV) resistance *N. glutinosa* (Lewis et al. 2005) or Tomato spotted wilt virus (TSWV) resistance from *N. alata* (Moon and Nicholson 2007). One drawback of this method is that the level of recombination when creating the first interspecific hybrid is low, and that a portion of *Nicotiana spp.* genome is introgressed along with the resistance gene, which in turn can cause undesirable characteristics for the farmers. It is thus interesting to introduce tobacco genes in order to bring resistances, without losing important commercial traits such as yield or quality.

Modern breeding lines are often male sterile. In tobacco, fertility restoration systems are not needed because the production of interest is the leaves rather than the seeds. Several sources of male sterility are available in tobacco and have been produced by the transfer of the *Nicotiana tabacum* genome into the cytoplasm of alien species of the same genus via backcrosses. As examples of sources of cytoplasmic male sterility (CMS), Wernsman and Rufty (Wernsman and Rufty 1987) describe how *N. suaveolens* is used as a source of CMS in hybrid burley seed production. Other cytoplasm were also used, including *N. bigelovii*, *N. plumbaginifolia*, *N. megalosiphon*, *N. undulate*, and *N. glauca*. The CMS cytoplasm can alter growth rate or cured leaf chemistry but does not necessarily impact yield or leaf quality (Lewis 2011); CMS is a good mechanism for plant variety protection and for hybrid production. In tobacco, hybrid vigor is not observed; the genetic effects are additive, meaning that commercial hybrids often exhibit an average phenotype of both parents.

Another important member of the *Nicotiana* genus for research and breeding in tobacco science is *Nicotiana africana*. It is used to develop haploid populations by crossing *N. tabacum* with *N. africana* pollen although anther culture can be also used to produce haploids in tobacco (Burk et al. 1979).

21.3 Genome

As the template for all biological information, the genome is the central and thus arguably most important ‘omics’ entity. Upon determining its sequence, one can analyze the chromosomal structure at the nucleotide level, find genes and determine their products. Of the many known challenges one faces during genome assembly,

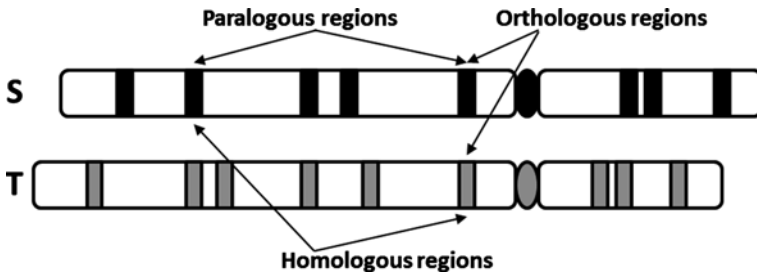


Fig. 21.2 The presence of 2 ancestral genomes in *N. tabacum* complicates its assembly. In addition to paralogous genes within each of the two individual genomes, they share orthologous genes between them. Their high level of similarity complicates read assignment during polyploid genome assembly

several are particularly pronounced for the *Nicotiana* species. The sheer size of the genome requires large amounts of data to be produced, stored and analyzed. In addition, the computational requirements, in particular the large amounts of memory needed, for assembling large genomes from short read sequences are immense and far exceed the abilities of current desktop computers. Furthermore, repetitive elements in the tobacco genome (Zimmerman and Goldberg 1977) can render sequencing intractable by current short read technologies: it may be impossible to align the fragments unambiguously if the sequencing reads do not span the entire length of a repeat region. Similarly, polyploidy may raise additional ambiguity when aligning or assigning reads (see Fig. 21.2), making the correct assembly even more of a challenge than with a ‘simple’ diploid species. These issues can be addressed respectively by using sequencing technologies producing longer reads, and by dividing the genome sequencing and assembly into smaller physical units, such as a minimum tiling path derived from physical map contigs.

21.3.1 *N. benthamiana* Genome Draft

N. benthamiana is a widely used model for plant-microbe biology. It is also responsive to virus-induced gene silencing (VIGS), thus facilitating the efficient functional study of plant genes. Native to Australia, it is an allotetraploid that formed from diploid parents from the *Sylvestres* and *Noctiflorae* sections (Knapp et al. 2004; Goodin et al. 2008). Its genome, estimated to be 3.5 Gb in length, is organized into 19 chromosomes. A draft genome sequence of *N. benthamiana* accession Nb-BTI was constructed at the Boyce Thompson Institute for Plant Research (BTI) (Gomez et al. 2012) and by The Commonwealth Scientific and Industrial Research Organisation (Naim et al. 2012). It was released in early 2012 through the Sol Genomics Network website (Bombarely et al. 2011). The assembly 0.4.2 of the *N. benthamiana* draft genome contains 141,339 scaffolds with an average length of 18.6 kb, thus covering 2.6 Gb and the assembly N50 length (the length of the smallest contig in the subset, whose combined length represents at least 50 % of the assembly) is 89,778 bp.

21.3.2 Tobacco Genome Initiative (*N. tabacum* and *N. benthamiana*)

Between 2001 and 2007, a project of the Tobacco Genome Initiative (TGI) was started in collaboration with Philip Morris USA to gather genetic information of *N. tabacum* by means of sequencing gene-rich regions of genomic DNA and cDNA libraries of the Hicks Broadleaf variety of tobacco. This project aimed at sequencing more than 90 % of *N. tabacum* genomic open reading frames and used the methyl-filtration method for genome complexity reduction. In addition, the BAC libraries necessary for a full-scale genome sequencing effort were constructed. More than 1.3 million Genome Survey Sequences (GSS) are now available in GenBank or through the NCSU portal. The assembly of these reads resulted in a fragmented genome with only 1-2 exons detected in the majority of contigs (Ivanov et al. 2010). While the *N. tabacum* genome data from the TGI is very informative, it is highly fragmented with the number of fragments corresponding to approximately 6 to 10 times the estimated number of genes. This fragmentation significantly reduces its applicability and use in plant breeding efforts. It is therefore crucial to continue the sequencing efforts and reach a state of at least near completion. On the side of transcriptome, a number of EST libraries from different organs (leaves, roots and flowers) and from different stresses (senescence, cold shock and viral infection) of *N. tabacum* and *N. benthamiana* were sequenced and released to GenBank. TGI data were leveraged to build a dense tobacco genetic map (Bindler et al. 2011), microsatellite marker kits for variety identification (Martin 2011) and a tobacco exon array (Martin et al. 2012).

21.3.3 PMI Physical Map Effort (*N. tabacum*)

Despite the progress of high throughput sequencing, BAC libraries remain an important component in genome sequencing projects. The construction of the physical map of a BAC library, which describes how the BACs are physically arranged, provides the basis for a BAC-by-BAC sequencing of the minimum tiling path and enables long-range scaffolding of assembly sequences and their anchoring on the chromosomes (see Fig. 21.3).

We have used this technique to characterize the allotetraploid genome of *N. tabacum* cv. Hicks Broadleaf, a breeding background of some flue-cured tobacco cultivars in use today. Four BAC libraries, comprising a total of 425,088 BAC clones and amounting to approximately 10.4 fold genome coverage were constructed and the novel sequence-based Whole Genome Profiling (WGPTM) technology¹ (van Oeveren et al. 2012) was used to obtain a tobacco physical map. These libraries were (1) a *Hind*III library consisting of 112,896 clones with an estimated average insert size of 100 kb, representing approximately 2.5x genome coverage; (2) a *Bam*HI library consisting of 146,304 clones with an estimated average insert size of 100 kb,

¹ The WGP technology is covered by patents and patent applications owned by Keygene N.V. WGP and KeyGene are (registered) trademarks of Keygene N.V.

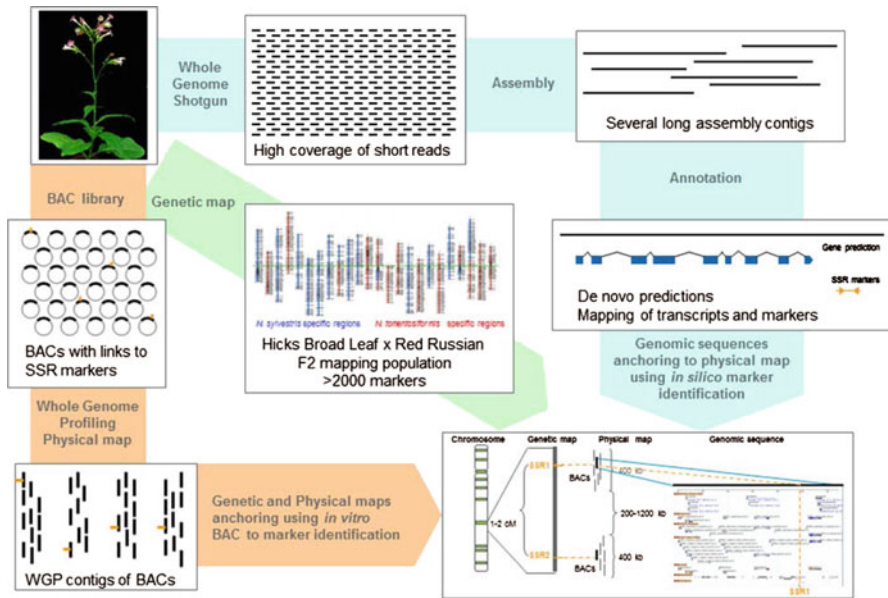


Fig. 21.3 The relationship between WGP physical map, the genetic map and the genome sequence of tobacco. This interrelationship allows genetic markers and BAC sequences to be linked: for a given marker, the relevant genome sequence can be quickly identified

representing approximately 3.4x genome coverage; (3) an *EcoRI* library consisting of 69,120 clones, with an estimated average insert size of 125 kb, representing approximately 1.9x genome coverage, and (4) a *HindIII* library consisting of 96,768 clones with an estimated average insert size of 125 kb, representing approximately 2.7x genome coverage. The obtained physical map contains 9,750 contigs of BACs with an average contig size of 462 kbp, and the calculated genome coverage equals the estimated tobacco genome size (Sierra et al. 2013a).

A novel method for determining the ancestral origin of the genome by annotation of WGP sequence tags was also applied (Sierra et al. 2013a), which agrees with the ancestral annotation available from the tobacco genetic map and may be used to investigate the evolution of homologous genome segments after polyploidization. The combination of the WGP physical mapping technology with tag profiling of ancestral lines represents a new generally applicable method to elucidate the ancestral origin of genome segments of polyploid species. Furthermore, polyploid plant biotechnology applications such as plant molecular farming for production of biologics are expected to be accelerated by the physical mapping of genes and their origins.

21.3.4 PMI Whole Genome Shotgun Effort (*N. tabacum*)

N. sylvestris and *N. tomentosiformis* originate from South America and exhibit for example different photoperiod sensitivity, flower color or diterpenoid production (Wagner 1991; Lin and Wagner 1994). Both are diploid species with 12 pairs of

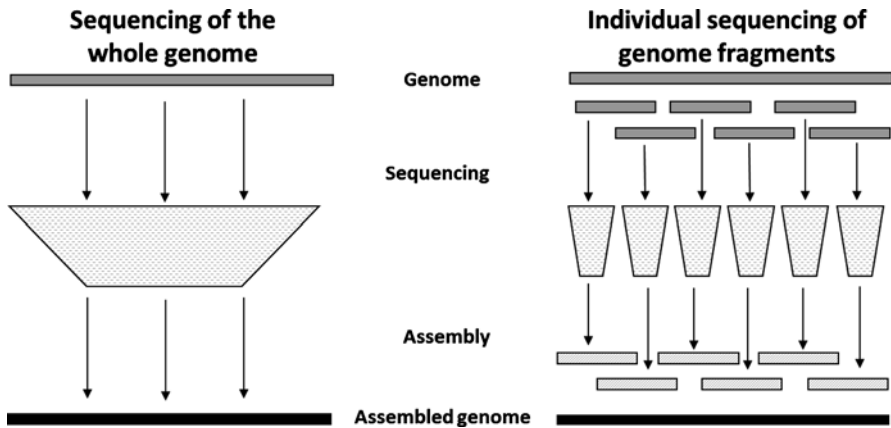


Fig. 21.4 Schematic representation of the workflow involved for genome sequencing

chromosomes and an estimated 1C genome size of 2.7 Gb, and are considered to be modern descendants of the respective maternal and paternal donors that formed the *N. tabacum* (common tobacco) about 200'000 years ago through interspecific hybridization (Leitch et al. 2008). *N. tabacum* is an amphidiploid (i.e., an allotetraploid that acts like a diploid) consisting of 24 chromosome pairs totaling 4.5 Gb (Fig. 21.4).

Contributing to the SOL-100 sequencing project (<http://solgenomics.net/organism/sol100/view>), which aims at sequencing 100 Solanaceae species, draft genomes for *N. sylvestris*, *N. tomentosiformis* and *N. tabacum* have been assembled by PMI from Illumina sequencing libraries and scaffolded using the tobacco physical map (Sierra et al. 2013b). The Max Planck Institute for Chemical Ecology in Jena, Germany, is working on the assembly of a *N. attenuata* draft genome.

21.3.5 Transcription Control

Transcription factors are essential in controlling the programming of the cell and so finding their sequences helps in identifying proteins involved in cell control and development. Rushton et al. (Rushton et al. 2008a, b) have searched for potential tobacco transcription factors within TGI data: their study reported 2513 sequences which can be classified into 64 families. This data is accessible through the TOBFAC database.

The binding site for the transcription factors is another important feature of the genome. Finding them allows downstream genes relevant to cell growth and development to be identified. Unfortunately, the prediction of the regions targeted by these transcription factors remains very challenging, especially on the fragmented genome assemblies, such as the TGI tobacco assembly, where the promoter sequence and the regulated gene may be located on separate contigs. It is therefore not possible to

rely on the presence of a coding gene downstream of the promoter to identify it. We developed an *ab initio* algorithm called dual-bagging that predicts promoter regions in the absence of the adjacent coding sequence. Our approach can be applied for analysis of many emerging genome sequences which suffer from fragmentation due to insufficient coverage or high repeat content.

21.3.6 Epigenomics

DNA methylation is an epigenetic phenomenon that controls gene expression and genome stability, making chromatin less accessible to transcription, transposition, DNA damage, and DNA repair. The accumulation of specific abiotic or biotic stress-induced transcripts in tobacco plants is associated with an active demethylation process (Wada et al. 2004; Choi and Sano 2007; Boyko and Kovalchuk 2011). Methylation polymorphism is widespread in *N. tabacum*, for example, among all of the *HapII-EcoRI/MspI-EcoRI* sites for 48 tobacco accessions, 49.3 % were methylated and 69.9 % of the allelic sites were polymorphic. A cluster pattern analysis showed that, among the accessions studied, geographic origin was closely related to methylation polymorphism; however, the genetic relationship was obscure. Methylation polymorphism may be useful as an epigenetic marker for certain narrow populations and cultivars affected by unique environments (Zhao et al. 2011). An investigation of two sibling tobacco cultivars, Yunyan85 and Yunyan87, and their two parents, K326 and Yunyan2 (Fu et al. 2012), demonstrated that 29 methylation-sensitive amplification polymorphism fragments exhibited methylation alteration in the four tobacco cultivars, thus supporting the hypothesis that methylation alteration of promoter regions could be responsible for the different phenotypes in tobacco. Thus, there is growing body of evidence that heritable epigenetic changes will play a more important role in the future breeding of tobacco.

21.4 Transcriptome

While the genome is defined as the nuclear genetic material of an individual organism, the transcriptome is potentially more difficult to delineate. Generally speaking it is the sum of all RNA sequences in a biological system. A system can be a cell, a tissue or an organism; in addition, the system can be in different states, for example subjected to a certain treatment. The transcriptome can include the primary transcript (heterogeneous nuclear RNA) as well as its processed derivatives, such as messenger RNA (mRNA), micro RNA (miRNA), small nuclear RNA (snRNA), and small interfering RNA (siRNA). Traditionally, mRNA expression has been the main focus of investigation, as it acts as a proxy to determine the expression level of proteins, which are the biochemical work horses of the cell. More recently, other types of RNA have been characterized in order to elucidate their function. As it is generally

faster and cheaper to obtain transcriptomic data than genomic information, transcriptomes have shown themselves to be useful in breeding studies for other polyploids (Bancroft et al. 2011).

Expression analysis is difficult in allopolyploids derived from closely related species, as orthologous genes from the two ancestor species complicate read assignment in addition to the already present homologous genes. Especially as the allotetraploidization event, which lead to the formation of tobacco, is estimated to be relatively recent, < 200,000 years ago according to (Clarkson et al. 2005), the unambiguous alignment becomes impossible for reads falling within non-divergent sites.

21.4.1 Expressed Sequence Tag Libraries

Before the advent of high throughput sequencing as it exists now, expressed sequence tags (ESTs) were a time and cost effective way of characterizing the transcriptome. An EST is a single read sequence from a cDNA library made by reversely transcribing the RNA from a sample. It represents the partial sequence of a transcript and thus can be used to identify genes in genomic DNA. The number of ESTs from a given sequence correlates with its frequency in the sample, and so the EST count can be used to obtain a rough estimate of a gene's expression level. As the technology necessary has been widely available for a long time, ESTs are an abundant and useful resource for gaining insight into the transcriptome.

Besides the efforts by individual research groups, several projects aimed specifically at generating EST data for *Nicotiana* species have provided a rich source of data. The TGI effort for example generated around 85k sequences for *N. tabacum* and 38k sequences for *N. benthamiana* (numbers in (Kole 2011)), and the European tobacco sequencing effort (ESTobacco) generated 46,546 sequences (Kole 2011). In addition, about 9,200 expressed sequences were generated by Matsuoka et al. (Matsuoka et al. 2004) (Table. 21.1).

In addition to the raw EST data, another useful resource is the Unigene repository: this resource maps ESTs to pseudo-genes and thus provides both reference transcriptome as well as the accompanying quantification of the transcripts contained within it. Unigene assemblies are available from the NCBI website; Solanaceae-specific assemblies can be found on the SOL website (<http://solgenomics.net/methods/unigene/index.pl>).

21.4.2 Microarray Analyses

Many phenotypes of importance in tobacco, such as morphology, leaf yield, heavy metal accumulation or nutrient deficiency, are thought to be transcriptionally controlled. Microarray technology is a suitable tool to study genetic variation and environmental effects with the objective to improve varieties of crops. For instance, EST-based tobacco microarrays have been used successfully in tobacco

Table 21.1 Public expressed sequence tag (EST) collection statistics

Species	Tissue	Number of ESTs
<i>N. attenuata</i>	Leaf	54
<i>N. attenuata</i>	Shoot	283
<i>N. attenuata</i>	Trichome	18
<i>N. benthamiana</i>	Leaf	16,969
<i>N. benthamiana</i>	Mixed	18,822
<i>N. benthamiana</i>	Not specified	13,316
<i>N. benthamiana</i>	Trichome	6,995
<i>N. glauca</i>	Leaf	8
<i>N. glauca</i> x <i>N. langsdorffii</i>	Leaf	2
<i>N. glauca</i> x <i>N. langsdorffii</i>	Not specified	3
<i>N. glauca</i> x <i>N. langsdorffii</i>	Tumor	3
<i>N. langsdorffii</i>	Leaf	13
<i>N. langsdorffii</i> x <i>N. sanderae</i>	Not specified	12,448
<i>N. megalosiphon</i>	Leaf	266
<i>N. megalosiphon</i>	Mixed	11
<i>N. sp.</i>	Flower	9
<i>N. suaveolens</i> x <i>N. tabacum</i>	Not specified	138
<i>N. sylvestris</i>	Leaf	8,583
<i>N. tabacum</i>	Anther	105
<i>N. tabacum</i>	Cell culture	543
<i>N. tabacum</i>	Embryo	1,647
<i>N. tabacum</i>	Flower	5,625
<i>N. tabacum</i>	Leaf	41,569
<i>N. tabacum</i>	Mixed	95,389
<i>N. tabacum</i>	Not specified	57,238
<i>N. tabacum</i>	Pollen	206
<i>N. tabacum</i>	Root	47,392
<i>N. tabacum</i>	Seed	4,361
<i>N. tabacum</i>	Seedling	33,169
<i>N. tabacum</i>	Shoot	21
<i>N. tabacum</i>	Trichome	6,296
<i>N. tabacum</i>	Whole	40,823

plant research (Edwards et al. 2010; Cui et al. 2011). Cui et al. compared the gene expression of selected 2,831 tobacco genes between the trichomes and the leaves with removed trichomes. Trichomes predominantly expressed genes involved in secondary metabolism, defense responses, and metabolic regulation (Cui et al. 2011). A Tobacco Expression Atlas (TobEA) (Edwards et al. 2010) was constructed through systematic measurements of gene expression across different tobacco samples. To achieve this, a custom built Affymetrix GeneChip™ was designed from the cDNA sequences originating from multiple tissues (seeds, roots, leaves, flowers, etc.) of several tobacco varieties. Nevertheless, none of these arrays were intended to cover genome-wide gene expression due to the lack of sufficient coverage of genomic sequences before the TGI data were released. An Affymetrix Exon Array was developed by Philip Morris International (PMI) based on the current genome and EST sequence data from the TGI to cover a large proportion of the tobacco gene space (Martin et al. 2012). The advantages of the exon array design include (i) the representation of the genes not

Table 21.2 Sequencing the transcriptomes of tobacco and its “ancestors”. Statistics of our in house transcriptome assemblies using the Trinity assembly pipeline (Grabherr et al. 2011)

Species	Tissue	Read pairs	Number of transcript sequences (unique gene models)	ORFs from transcript sequences (unique gene models)
<i>N. sylvestris</i>	Root	86,993,889	195,925 (147,650)	25,153 (14,334)
<i>N. sylvestris</i>	Leaf	56,623,372	145,491 (92,930)	23,894 (9,956)
<i>N. sylvestris</i>	Flower	108,436,707	170,185 (117,471)	23,361 (11,659)
<i>N. tomentosiformis</i>	Root	80,974,445	246,733 (195,892)	29,892 (19,001)
<i>N. tomentosiformis</i>	Leaf	68,790,307	146,822 (94,772)	22,566 (9,265)
<i>N. tabacum</i>	Root	75,646,377	338,099 (188,859)	41,999 (19,952)
<i>N. tabacum</i>	Leaf	90,682,315	282,992 (127,554)	30,974 (9,514)
<i>N. tabacum</i>	Flower	128,576,512	349,045 (159,542)	34,390 (10,765)

yet found in the currently available EST libraries, (ii) the ability to investigate alternative splicing in tobacco plant, (iii) the equal probe coverage of each exon of the gene. The Tobacco Exon Array has been used in studies of cadmium accumulation in leaves. The experiments described in this work are available through GeneVestigator (Hruz et al. 2008) plant database and SGN web sites.

21.4.3 RNA-Seq Effort

A sequencing effort was launched by PMI in order to obtain the transcriptomes of *N. tabacum* and the descendants of its parent species, *N. sylvestris* and *N. tomentosiformis*. The transcriptome will allow the identification of genes and analyze their expression; thus we hope to be able to establish changes in regulation which have ensued following species hybridization, such as the deletion or silencing of duplicate genes. By using different tissues, the presence and expression of the various genes in these tissues will be able to be gauged. The “meta-transcriptome”, that is the sum of all the transcripts in all of the tissues sampled will provide meaningful annotation for the PMI genome assembly effort, allowing genes to be identified or confirmed (Table. 21.2).

21.4.4 Small RNA Data

In most eukaryotes, gene expression can be regulated by transcriptional and post-transcriptional silencing mechanisms involving small RNAs. The comparative sequencing of plant small RNA database (Mahalingam and Meyers 2010) (available at <http://smallrna.udel.edu>), contains microRNA (miRNA) and short interfering RNA (siRNA) sequences obtained from a wide variety of plants, as well as tools for analysis and comparison. For *N. tabacum*, 1,608,893 distinct sequences are available for leaves, 2,966,609 for flowers and 1,024,045 for pod.

21.5 Proteome

Advances in sequencing technologies have yielded strong progress for those ‘omics’ fields which involve nucleic acid sequences. The availability of genomic data has allowed the protein complement of some Nicotiana species to be studied by bioinformatics techniques. In addition, spectrometric methods for studying protein sequences have advanced steadily and been able to benefit from the complementary genomic information.

21.5.1 Protein Resources and Prediction From the Genome

The best sources of protein information are the expert-annotated databases. For some well-studied model organism plants like *A. thaliana*, there are species specific databases which are supported by a large number of researchers who contribute to its curation. Although, to the best of our knowledge, there is no such resource available for Nicotiana species, much information is available in the ‘general’ databases. Bioinformatics techniques for predicting proteins in genomes, such as mutual best blast hit search or the Eficaz2 tool (Tian et al. 2004; Arakaki et al. 2009), can be used to annotate putative protein sequences predicted from genomic data.

21.5.2 Proteomics Characterization Using Tandem Mass Spectrometry

Mass spectrometry (MS) is a technique which allows constituents in a sample to be separated and identified based on the ratio of their mass and charge; this information can be used to identify the peptides from which they were derived. For very high quality data the peptide sequence can be deduced directly from the spectra (Seidler et al. 2010); more usually however, the spectra are matched to an *in silico* digest of a reference protein database, as is done by algorithms such as SEQUEST (Ducret et al. 1998), Mascot (Perkins et al. 1999), X!Tandem (Craig and Beavis 2004) and OMSSA (Geer et al. 2004). For Nicotiana species, studies using this method have been limited by the availability of genomic data for reference (Millar et al. 2009). There are no sequence databases specifically tailored to Nicotiana species, but numerous resources do contain Nicotiana protein sequences, or sequences for species close enough to allow the identification of peptides. A number of studies have characterized the proteome of tobacco (Dani et al. 2005) or tobacco BY-2 cells (Duby et al. 2010). Protein assignment is generally achieved by matching spectra against a database derived from a variety of different plants, for example one study (Duby et al. 2010) used a diverse set of reference proteins from Nicotiana, Solanaceae as well as distantly related species such as Arabidopsis.

As for dedicated resources for collecting proteomics data, there are currently only few available, particularly for plant data. The ProMEX database, however, is a promising project for the future: it is a proteomics database containing plant peptide fragments identified from proteomics spectral data (Hummel et al. 2007; Wienkoop et al. 2012). It contains peptide hits from a number of plant species including Solanaceae and although few *Nicotiana*-specific peptides are identified in the database, the future release of more genomic sequences for reference will hopefully see the coverage extend.

A more recent application of tandem MS data is the annotation of genomic sequences using spectra. Instead of using selected proteins only, proteogenomics (Jaffe et al. 2004) reverses this process by generating all possible peptides from all possible ORFs identified in genomic material and matching the shotgun proteomics data against it, thus confirming actual ORFs as coding regions in absence of gene models. While there have been, to the best of our knowledge, no studies of this kind on tobacco which have been published in scientific literature, the availability of unannotated genomic data make this an appealing technique for future projects.

21.6 Metabolome

The small molecules complement of plants (Eich 2008) and the metabolic pathways by which they are produced have become a recent focal point of interest. Around 4,200 constituents of tobacco have been identified (Rodgman and Perfetti 2008) and the pathways by which they are generated are of obvious interest. These network maps act not only as a resource for qualitative data, but they also form the basis for quantitative metabolic studies such as flux balance analysis of metabolic flux analysis. Using genomic and proteomic data, the presence of enzymes in an organism can be inferred, and by extension so can the corresponding metabolic pathway.

21.6.1 Metabolic Network Resources

The KEGG database (Kanehisa and Goto 2000; Kanehisa et al. 2012) provides metabolic information on a number of hierarchical levels: it lists reactions, which can be grouped into pathways and which in turn can be finally assembled into whole network maps. Given experimental evidence for enzyme activity as well as predicted enzyme function, species pathways can be reconstructed, though not in a fully automated manner. The Pathway Tools software suite (Karp et al. 2002a) provides a method to create organism-specific metabolic databases. Given a set of available enzymes, this software can identify likely pathways in an organism using a reference database. The software provides a server for publishing the data online, and the SOL Genomics Network website provides several such databases for Solanaceae species, including tobacco. The quality of such databases is determined by the reliability of

the annotation of enzyme complement of a plant. In many cases, protein function prediction methods have to be used to determine the likely occurrence of a given function in the genome. Given the enzyme complement of the cell it is now possible to identify candidate pathways from a reference database; besides the default reference database, MetaCyc (Karp et al. 2000; Karp et al. 2002b), a specialized plant pathway database, PlantCyc (Zhang et al. 2010), is available.

21.6.2 Application of Metabolomics Information

The fields of metabolomics and metabonomics (Nicholson et al. 1999) concern themselves with the study of the metabolic complement of the cell and the rules which govern it. Numerous studies on the tobacco metabolism have been performed, including metabolic fingerprinting (Choi et al. 2004), which have been performed using gas and liquid phase chromatography in conjunction with mass spectrometry (Li et al. 2011a, b). Metabolic resources can act as a resource of information for researchers, providing candidate molecules and their properties for quick molecular identification. Other experiments have investigated the link between metabolomics and transcriptomics (Misra et al. 2010): by finding genes co-expressed with those responsible for certain enzymatic steps in pathways of interest, one can identify probable candidates for missing steps. Another area of interest is the comparison of the metabolome of various species of the *Nicotiana* genus or strains of the various species. Such comparative metabolomics has been performed for transgenic varieties of *N. tabacum* var. *Xanthi* (Broeckling et al. 2012). Using enzyme data, metabolic maps can be created from the genome sequence, allowing comparative analysis *in silico*.

With the advent of large scale metabolic maps—the most extensive map for a plant species was produced for *Arabidopsis thaliana* (Poolman et al. 2009) – quantitative modeling of the plant metabolome has become possible. Metabolic flux analysis (MFA) for example, requires a map of the metabolism in order to trace the flow of metabolites; using a map of the core metabolism, MFA has been performed for tobacco (Ettenhuber et al. 2005). These analyses require a stoichiometrically complete map (the number and identity of all metabolites participating in a reaction have to be defined) of the metabolic network, which can be assembled from the data contained in the Pathway Tools and KEGG platforms.

21.7 Summary

The extent of the ‘Omics’ resources for the *Nicotiana* genus of the Solanaceae is still comparatively limited compared to those of other plants, despite the interest in and importance of the *Nicotiana* genus. While the genetics is well characterized, several genomes are currently still being sequenced and many resources are being

developed which will become available in the near future. The genome sequences will provide more detail on the differences between the *Nicotiana* species, in particular the varieties of tobacco, and help resolve misclassifications. This in turn may lead to improved breeding strategies, by more accurately identifying the genetic basis of desirable traits. It will also improve our understanding of the processes underlying polyploid evolution and advance the use of *Nicotiana* species as model organisms. Proteomics and metabolomics are well established methods which are currently limited by the availability of genomic data; a deluge of knowledge can be expected to be produced upon publication of *Nicotiana* genomes in the foreseeable future.

References

- Arakaki AK, Huang Y, Skolnick J (2009) EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. *BMC Bioinformatics* 10:107
- Bancroft I, Morgan C, Fraser F et al (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. *Nat Biotechnol* 29:762–766
- Bindler G, van der Hoeven R, Gunduz I et al (2007) A microsatellite marker based linkage map of tobacco. *Theor Appl Genet* 114:341–349
- Bindler G, Plieske J, Bakaheer N et al (2011) A high density genetic map of tobacco (*Nicotiana tabacum* L.) obtained from large scale microsatellite marker development. *Theor Appl Genet* 123:219–230
- Bombarely A, Menda N, Teclé IY et al (2011) The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. *Nucleic Acids Res* 39:1149–1155
- Boyko A, Kovalchuk I (2011) Genetic and epigenetic effects of plant-pathogen interactions: an evolutionary perspective. *Mol Plant* 4:1014–1023
- Broeckling CD, Li K-G, Xie D-Y (2012) Comparative Metabolomics of Transgenic Tobacco Plants (*Nicotiana tabacum* var. *Xanthi*) Reveals Differential Effects of Engineered Complete and Incomplete Flavonoid Pathways on the Metabolome. In: Çiftçi YÖ (ed) *Transgenic Plants—Advances and Limitations*. InTech
- Burk L, Gerstel D, Wernsman E (1979) Maternal haploids of *Nicotiana tabacum* L. from seed. *Science* 206:585–585
- Chase MW, Knapp S, Cox AV et al (2003) Molecular systematics, GISH and the origin of hybrid taxa in *Nicotiana* (Solanaceae). *Ann Bot* 92:107–127
- Choi CS, Sano H (2007) Abiotic-stress induces demethylation and transcriptional activation of a gene encoding a glycerophosphodiesterase-like protein in tobacco plants. *Mol Genet Genomics* 277:589–600
- Choi HK, Choi YH, Verberne M et al (2004) Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique. *Phytochemistry* 65:857–864
- Clarkson JJ, Lim KY, Kovarik A et al (2005) Long-term genome diploidization in allopolyploid *Nicotiana* section *Repandae* (Solanaceae). *New Phytol* 168:241–252
- Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20:1466–1467
- Cui H, Zhang ST, Yang HJ et al (2011) Gene expression profile analysis of tobacco leaf trichomes. *BMC Plant Biol* 11:76
- Dani V, Simon WJ, Duranti M, Croy RR (2005) Changes in the tobacco leaf apoplast proteome in response to salt stress. *Proteomics* 5:737–745
- Davalieva K, Maleva I, Filiposki K et al (2010) Genetic Variability of Macedonian tobacco varieties determined by microsatellite marker analysis. *Diversity* 2:439–449

- Denduangboripant J, Piteekan T, Nantharat M (2010) Genetic polymorphism between tobacco cultivar-groups revealed by amplified fragment length polymorphism analysis. *J Agricul Sci* 2:41
- Duby G, Degand H, Faber AM, Boutry M (2010) The proteome complement of *Nicotiana tabacum* Bright-Yellow-2 culture cells. *Proteomics* 10:2545–2550
- Ducret A, Van Oostveen I, Eng JK et al (1998) High throughput protein characterization by automated reverse-phase chromatography/electrospray tandem mass spectrometry. *Protein Sci* 7:706–719
- Edwards KD, Bombarely A, Story GW et al (2010) TobEA: an atlas of tobacco gene expression from seed to senescence. *BMC Genomics* 11:142
- Eich E (2008) Solanaceae and Convolvulaceae: Secondary Metabolites: Biosynthesis, Chemotaxonomy, Biological and Economic Significance (A Handbook). Springer
- Ettenhuber C, Radykewicz T, Kofer W et al (2005) Metabolic flux analysis in complex isotopolog space. Recycling of glucose in tobacco plants. *Phytochemistry* 66:323–335
- Farhi M, Marhevka E, Ben-Ari J et al (2011) Generation of the potent anti-malarial drug artemisinin in tobacco. *Nat Biotechnol* 29:1072–1074
- Fricano A, Bakaher N, Del CM (2012) Molecular diversity, population structure, and linkage disequilibrium in a worldwide collection of tobacco (*Nicotiana tabacum* L.) germplasm. *BMC Genet* 13:18
- Fu SL, Tang ZX, Liu L et al (2012) Variation of genomic DNA methylation in the nitrate reductase gene of sibling tobacco (*Nicotiana tabacum*) cultivars. *Genet Mol Res* 11:1169–1177
- Geelen DN, Inze DG (2001) A bright future for the bright yellow-2 cell culture. *Plant Physiol* 127:1375–1379
- Geer LY, Markey SP, Kowalak JA et al (2004) Open mass spectrometry search algorithm. *J Proteome Res* 3:958–964
- Gomez AB, Vrebalov J, Moffett P et al (2012) A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant-microbe biology research. *Mol Plant Microbe Interact* 25:1523–30
- Goodin MM, Zaitlin D, Naidu RA, Lommel SA (2008) *Nicotiana benthamiana*: its history and future as a model for plant-pathogen interactions. *Mol Plant Microbe Interact* 21:1015–1026
- Goodspeed TH (1954) The genus *Nicotiana*. *Chronica Botanica* 16:102–135
- Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652
- Hruz T, Laule O, Szabo G et al (2008) Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics* 2008
- Hummel J, Niemann M, Wienkoop S et al (2007) ProMEX: a mass spectral reference database for proteins and protein phosphorylation sites. *BMC Bioinformatics* 8:216
- Ivanov NV, Sierro N, Gadani F, Peitsch MC (2010) Current State of Tobacco Genome Sequencing. In *Plant and Animal Genome XVIII Conference*
- Jaffe JD, Berg HC, Church GM (2004) Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* 4:59–77
- Julio E, Denoyes-Rothan B, Verrier JL, Dorlhac de Borne F (2006) Detection of QTLs linked to leaf and smoke properties in *Nicotiana tabacum* based on a study of 114 recombinant inbred lines. *Mol Breeding* 18:69–91
- Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28:27–30
- Kanehisa M, Goto S, Sato Y et al (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 40:109–114
- Karp PD, Paley S, Romero P (2002a) The Pathway Tools software. *Bioinformatics* 18(1):225–232
- Karp PD, Riley M, Paley SM, Pellegrini-Toole A (2002b) The MetaCyc Database. *Nucleic Acids Res* 30:59–61
- Karp PD, Riley M, Saier M et al (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res* 28:56–59

- Kelly LJ, Leitch AR, Clarkson JJ et al (2010) Intragenic recombination events and evidence for hybrid speciation in *Nicotiana* (Solanaceae). *Mol Biol Evol* 27:781–799
- Khan MQ, Narayan R (2007) Phylogenetic diversity and relationships among species of genus *Nicotiana* using RAPDs analysis. *Afr J Biotechnol* 6
- Knapp S, Chase MW, Clarkson JJ (2004) Nomenclatural changes and a new sectional classification in *Nicotiana* (Solanaceae). *Taxon* 53:73–82
- Kole C (ed) (2011) *Wild Crop Relatives: Genomic and Breeding Resources*. Springer-Verlag
- Leitch IJ, Hanson L, Lim KY et al (2008) The ups and downs of genome size evolution in polyploid species of *Nicotiana* (Solanaceae). *Ann Bot* 101:805–814
- Lewis RS (2011) *Nicotiana*. *Wild crop relatives: genomic and breeding resources*. 185–208
- Lewis RS, Kernodle SP (2009) A method for accelerated trait conversion in plant breeding. *Theor Appl Genet* 118:1499–1508
- Lewis RS, Milla SR, Levin JS (2005) Molecular and genetic characterization of *Nicotiana glutinosa* L. chromosome segments in tobacco mosaic virus-resistant tobacco accessions. *Crop sci* 45:2355–2362
- Lin Y, Wagner G (1994) Surface disposition and stability of pest-interactive, trichome-exuded diterpenes and sucrose esters of tobacco. *J chem ecol* 20:1907–1921
- Li Q, Zhao C, Li Y et al (2011a) Liquid chromatography/mass spectrometry-based metabolic profiling to elucidate chemical differences of tobacco leaves between Zimbabwe and China. *J Sep Sci* 34:119–126
- Li Y, Pang T, Wang X et al (2011b) Gas chromatography-mass spectrometric method for metabolic profiling of tobacco leaves. *J Sep Sci* 34:1447–1454
- Ling HY, Edwards AM, Gantier MP et al (2012) An interspecific *Nicotiana* hybrid as a useful and cost-effective platform for production of animal vaccines. *PLoS One* 7:35688
- Liu XZ, ShenHe C, Yang YM, ZHANG HY (2009) Genetic diversity among flue-cured tobacco cultivars on the basis of AFLP markers. *Czech J Genet Plant Breed* 45:155–159
- Mahalingam G, Meyers BC (2010) Computational methods for comparative analysis of plant small RNAs. *Methods Mol Biol* 592:163–181
- Martin F (2011) An application of kernel methods to variety identification based on SSR markers genetic fingerprinting. *BMC Bioinforma* 12:177
- Martin F, Bovet L, Cordier A et al (2012) Design of a tobacco exon array with application to investigate the differential cadmium accumulation property in two tobacco varieties. *BMC Genetics* 13:674
- Matsuoka K, Demura T, Galis I et al (2004) A comprehensive gene expression analysis toward the understanding of growth and differentiation of tobacco BY-2 cells. *Plant Cell Physiol* 45:1280–1289
- Milla SR, Levin JS, Lewis RS, Ruffy RC (2005) RFLP and scar markers linked to an introgressed gene conditioning resistance to D.b. Adam. in *Tobacco*. *Crop Sci* 45:2346–2354
- Millar DJ, Whitelegge JP, Bindschedler LV et al (2009) The cell wall and secretory proteome of a tobacco cell line synthesising secondary wall. *Proteomics* 9:2355–2372
- Misra P, Pandey A, Tiwari M et al (2010) Modulation of transcriptome and metabolome of tobacco by *Arabidopsis* transcription factor, AtMYB12, leads to insect resistance. *Plant Physiol* 152:2258–2268
- Moon H, Nicholson J (2007) AFLP and SCAR markers linked to tomato spotted wilt virus resistance in tobacco. *Crop science* 47:1887–1894
- Moon HS, Nifong JM, Nicholson JS et al (2009) Microsatellite-based Analysis Of Tobacco (L.) Genetic Resources. *Crop Sci* 49:2149–2159
- Murad L, Lim KY, Christopolulou V et al (2002) The origin of tobacco's T genome is traced to a particular lineage within *Nicotiana tomentosiformis* (Solanaceae). *Am J Bot* 89:921–928
- Nagata T, Hasewa S, Inzé D (eds) (2004) *Tobacco BY-2 Cells*. Springer
- Naim F, Nakasugi K, Crowhurst RN et al (2012) Advanced engineering of lipid metabolism in *Nicotiana benthamiana* using a draft genome and the V2 viral silencing-suppressor protein. *PLoS One* 7:e52717

- Nicholson J, Lewis R, van der Hoeven R et al (2009) Microsatellite-based analysis of Tobacco (*L.*) genetic resources. *Crop sci* 49:2149–2159
- Nicholson JK, Lindon JC, Holmes E (1999) ‘Metabonomics’: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29:1181–1189
- Perkins DN, Pappin DJC, Creasy DM, Cottrell JS (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20:3551–3567
- Poolman MG, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of *Arabidopsis* and some of its properties. *Plant Physiol* 151:1570–1581
- Ren N, Timko MP (2001) AFLP analysis of genetic polymorphism and evolutionary relationships among cultivated and wild *Nicotiana* species. *Genome* 44:559–571
- Renny-Byfield S, Chester M, Kovarik A et al (2011) Next generation sequencing reveals genome downsizing in allotetraploid *Nicotiana tabacum*, predominantly through the elimination of paternally derived repetitive DNAs. *Mol Biol Evol* 28:2843–2854
- Rodgman A, Perfetti TA (2008) The chemical components of tobacco and tobacco smoke. CRC
- Rushton PJ, Bokowiec MT, Han S et al (2008a) Tobacco transcription factors: novel insights into transcriptional regulation in the Solanaceae. *Plant Physiol* 147:280–295
- Rushton PJ, Bokowiec MT, Laudeman TW et al (2008b) TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics* 9:53
- Sarala K, Rao RVS (2008) Genetic diversity in Indian FCV and burley tobacco cultivars. *J genet* 87:159–163
- Seidler J, Zinn N, Boehm ME, Lehmann WD (2010) De novo sequencing of peptides by MS/MS. *Proteomics* 10:634–649
- Sierro N, Van Oeveren J, van Eijk MJ et al (2013a) Whole genome profiling physical map and ancestral annotation of tobacco hicks broadleaf. *Plant J* 75:880–889
- Sierro N, Batteny JND, Ouali S et al (2013b) Reference genomes and transcriptomes of *Nicotiana sylvestris* and *Nicotiana tomentosiformis*. *Genome Biology* 14:R60
- The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485:635–641
- Tian W, Arakaki AK, Skolnick J (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. *Nucleic Acids Res* 32:6226–6239
- van Oeveren J, de Ruyter M, Jesse T et al (2011) Sequence-based physical mapping of complex genomes by whole genome profiling. *Genome Res* 21:618–625
- Vontimitta V, Daneshmandi DA, Steede T et al (2010) Analysis of a *Nicotiana tabacum* L. genomic region controlling two leaf surface chemistry traits. *J Agric Food Chem* 58:294–300
- Vontimitta V, Lewis RS (2012) Mapping of quantitative trait loci affecting resistance to *Phytophthora nicotianae* in tobacco (*Nicotiana tabacum* L.) line Beinhart-1000. *Molecular Breeding*. 1–10
- Wada Y, Miyamoto K, Kusano T, Sano H (2004) Association between up-regulation of stress-responsive genes and hypomethylation of genomic DNA in tobacco plants. *Mol Genet Genomics* 271:658–666
- Wagner GJ (1991) Secreting glandular trichomes: more than just hairs. *Plant Physiol* 96:675
- Wernsman EA, Rufty RC (1987) Tobacco. In: Fehr WR (ed) Principles of cultivar development Volume 2 Crop species. Macmillan publishing company
- Wienkoop S, Staudinger C, Hoehenwarter W et al (2012) ProMEX—a mass spectral reference database for plant proteomics. *Front Plant Sci* 3:125
- Wu F, Mueller LA, Crouzillat D et al (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. *Genetics* 174:1407–1420
- Wu F, Eannetta NT, Xu Y et al (2010) COSII genetic maps of two diploid *Nicotiana* species provide a detailed picture of synteny with tomato and insights into chromosome evolution in tetraploid *N. tabacum*. *Theor Appl Genet* 120:809–827

- Zhang P, Dreher K, Karthikeyan A et al (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol* 153:1479–1491
- Zhao JH, Zhang JS, Wang Y et al (2011) DNA methylation polymorphism in flue-cured tobacco and candidate markers for tobacco mosaic virus resistance. *J Zhejiang Univ Sci B* 12:935–942
- Zimmerman JL, Goldberg RB (1977) DNA sequence organization in the genome of *Nicotiana tabacum*. *Chromosoma* 59:227–252

Chapter 22

Mining SNPs and Linkage Analysis in *Cynara cardunculus*

Sergio Lanteri, Alberto Acquadro, Davide Scaglione and Ezio Portis

Contents

22.1 The <i>Cynara cardunculus</i> Complex	534
22.2 Uses of Globe Artichoke and Cardoon other than for Human Food	536
22.3 Linkage Analyses: State of the Art	537
22.4 SNP Mining	541
22.4.1 Genomic SNP Mining	543
22.4.2 Transcriptomic SNP Mining	546
22.4.3 Conclusions	551
References	553

Abstract *Cynara cardunculus* L., a member of the *Asteraceae* family, is a diploid ($2n = 34$) outcrossing perennial species native to the Mediterranean basin. It includes globe artichoke (var. *scolymus* L.), which today is grown as vegetable all over the world, cultivated cardoon (var. *altilis* DC), locally grown in Southern European countries, and their progenitor wild cardoon (var. *sylvestris* Lam). The species is also a valuable source of pharmaceutical compounds, and is exploitable for the production of lignocellulosic biomass as well as oil from seed, the latter being suitable for both edible and bio-fuel end-uses.

By crossing a non spiny globe artichoke genotype (female parent) with selected genotypes of the three botanical taxa, we generated three F1 segregating progenies from which genetic maps, based on the two-way pseudo test cross strategy, have been developed. From the globe artichoke and cultivated cardoon genetic maps a reference SSR-based consensus map was constructed, which consists of 227 loci (217 SSRs and ten SNPs) assembled into 17 major linkage groups. To further saturate the *C. cardunculus* maps we recently applied NGS (next generation sequencing) technologies for mining a wide set of SNPs (single nucleotide polymorphism). Based on Illumina sequencing of gDNA RAD (restriction associated DNA) tags of three mapping parents (e.g. non spiny globe artichoke, cultivated and wild cardoon), we generated ~ 19,000

S. Lanteri (✉) · A. Acquadro · D. Scaglione · E. Portis
University of Torino, DISAFA, Via Leonardo da Vinci 44, 10095,
Grugliasco TO, Italy
e-mail: sergio.lanteri@unito.it

genomic contigs (mean 312 bp) and $\sim 17,000$ SNPs (density 1/139 bp). Side by side, the transcriptome of the same mapping parents was sequenced by using a 454 platform, and raw data *de novo* assembled and annotated to generate the first reference transcriptome of the species (38,726 unigenes, 32.7 Mbp).

The 454 reads, together with Illumina paired ends (PEs) from further eight *C. cardunculus* genotypes were aligned on the reference contig set, and $\sim 195,000$ SNPs were called (density 1/169 bp in coding regions). The two workflows led to produce a massive set of SNPs in *C. cardunculus*, and made possible create an extensive gene catalogue as a valuable resource for upcoming genomic and genetic studies.

22.1 The *Cynara cardunculus* Complex

Cynara cardunculus L. is native to the Mediterranean Basin and includes three botanical taxa: the globe artichoke (var. *scolymus*), the cultivated cardoon (var. *altilis*) and the wild cardoon [var. *sylvestris* (Lamk) Fiori]. The three forms are fully cross-compatible with one another, and form fertile hybrids (Basnizki and Zohary 1994). Reproductive barriers separate the *C. cardunculus* complex from the other *Cynara* species (Rottenberg et al. 1996). The crosses between *C. cardunculus* and the wild species *C. syriaca*, *C. algarbiensis*, *C. baetica* and *C. humilis* do all produce few seeds, although the hybrids are generally sterile; the wild species are therefore regarded as members of the secondary wild gene pool of *C. cardunculus* (Rottenberg and Zohary 2005). On both morphological (Wiklund 1992) and cytogenetic (Rottenberg et al. 1996) grounds, the closest of the wild species to the cultivated complex is *C. syriaca*. The monophyly and evolution of the *Cynara* spp. have been investigated by sequence comparisons between various ITS (internal transcribed spacer) regions (Robba et al. 2005 leading to the suggestion that the *cardunculus* complex is more differentiated and evolved than the other wild species.

Molecular (Lanteri et al. 2004; Acquadro et al. 2005), cytogenetic and isozyme (Rottenberg et al. 1996) studies have confirmed that wild cardoon is the ancestor of both the domesticated globe artichoke and cultivated cardoon, which evolved independently under the influence of distinct anthropogenic selection criteria. The earliest report of the presence of *C. cardunculus* in Sicily and Greece dates back to Theophrastus (371–287 BCE), while in 77 CE, the Roman naturalist Pliny the Elder mentioned its use for medicinal purposes; however, little is known either of the process of domestication or the subsequent diversification of the two taxa. It has been assumed that before globe artichoke was selected, cardoon was cultivated for its fleshy stems and roots, which were considered a delicacy by the ancient Greeks and Romans (Portis et al. 2005a, b). On the other hand, the best guess is that the globe artichoke was domesticated and transformed into the plant that we know today, most probably between 800 and 1500 CE in family or monastery gardens. Recently, by assessing the AFLP pattern of genetic diversity of a collection of Sicilian globe artichoke landraces, which have been cultivated for a number of centuries by local

farmers, one landrace was identified which appears to represent an early stage of the domestication process, suggesting Sicily as one of the possible centre of globe artichoke domestication (Mauro et al. 2008).

Globe artichoke contributes significantly to the Mediterranean agricultural economy, with an annual production of about 750 metric tons (MT) from over 80,000 ha of cultivated land and with an annual turnover exceeding US \$ 500 million. Italy is the leading world producer (475 MT/year, FAOSTAT 2011; <http://faostat.fao.org/>), followed by Egypt and Spain. Globe artichoke cultivation is increasing in South America and the United States, and more recently also in China. The prime globe artichoke product consists of the immature inflorescence (heads of capitula), which can be consumed in fresh, canned or frozen form. Each plant produces a number of capitula, the largest of which (the main capitulum) merges from the apex of the central stem, while the smaller ones are produced on lateral branches.

Italy has the richest globe artichoke primary cultivated “gene pool” and harbours many distinct clonal varietal groups, best adapted to local environments. On the basis of harvest time, varietal types can be classified as early, producing heads from autumn to spring, and late, producing heads from early to late spring. On the basis of capitulum characters, cultivated germplasm has been classified into four main groups: (1) the Spinosi group, containing types with long sharp spines on bracts and leaves; (2) the Violetti group, with medium-sized, violet-coloured and less spiny heads; (3) the Romaneschi group, with spherical or subspherical non-spiny heads; (4) the Catanesi group, with relatively small, elongated and non-spiny heads. The classification based on head is in consistent agreement with the one obtained by assessing the AFLP genetic variation in a wide collection of 84 varietal types grown worldwide, indicating that the cultivated morphotypes play an important role in determining variation within the cultivated globe artichoke germplasm (Lanteri et al. 2004). Although in recent years some seed (achenes)-propagated varieties have been introduced, but vegetative propagation, by means of basal and lateral offshoots (either semi-dormant or actively growing), or stump pieces, has been adopted for centuries, and it is still largely prevalent in most of the varietal types and local landraces. Due to the limited selection adopted by farmers on the mother plants used for vegetative propagation, as well as mutations occurred over time, the populations at present in cultivation are multiclonal and characterized by a wide range of within population genetic variation (Lanteri et al. 2001; Portis et al. 2005c).

The cultivated cardoon is usually raised from seed and handled as annual crop; its cultivation is much less widespread than that of the globe artichoke and the crop remains of regional importance in Spain, Italy and the south of France, where it is used in traditional dishes. The edible parts of the plant are the fleshy stems which are typically collected in late autumn-early winter and often, before collection are tied together, wrapped in straw, and/or buried for about three weeks in order to accentuate the flavour. A study based on SSR and AFLP profiling of the most widely grown Italian and Spanish local varieties showed that they form two separate gene-pools and that a considerable level of within variety variation is present (Portis et al. 2005b).

The wild cardoon is a robust thistle distributed over the west and central part of the Mediterranean basin (Portugal to west Turkey) as well as Canary Islands; in post

Columbian time it colonized some part of the New World and has spread as a weed in Argentina and California (Marushia and Holt 2006). Its flowers have been used for centuries in the Iberian Peninsula for manufacturing of ovine and caprine milk cheese (Sousa and Malcata 1996; Barbagallo et al. 2007) and its small and thorny capitula are sometimes sold in local markets in Sicily (Ierna and Mauromicale 2010).

22.2 Uses of Globe Artichoke and Cardoon other than for Human Food

C. cardunculus has long been known to represent a valuable source of biopharmaceutical compounds (Slanina et al. 1993; Wagenbreth 1996; Sevcikova et al. 2002; Wang et al. 2003). Roots and rhizomes, used also for brew or infusion, provide a source of inulin, a demonstrated enhancer of the human intestinal flora, while leaves and heads represent one of the richest natural source of compounds originating from the metabolism of phenylpropanoids, with caffeoylquinic acids and flavonoids as major components. *C. cardunculus* extracts influence glucose and lipid metabolism (Blumenthal et al. 2000) and were reported to be effective in increasing the feeling of satiety in overweight subjects (Rondanelli et al. 2011); in various pharmacological test systems it has been demonstrated that they (i) protect proteins lipids and DNA from oxidative damage from free radicals (Gebhardt 1997; Brown and Rice-Evans 1998; Perez-Garcia et al. 2000), (ii) inhibit cholesterol biosynthesis and contribute to the prevention of atherosclerosis and other vascular disorders (Kraft 1997; Brown and Rice-Evans 1998; Gebhardt 1998; Pittlern and Ernst 1998; Matsui et al. 2006; Bundy et al. 2008). Furthermore, it has been demonstrated that *C. cardunculus* extracts inhibit HIV integrase, a key player in HIV replication and its insertion into host DNA (McDougall et al. 1998; Slanina et al. 2001), possess apoptotic properties (Miccadei et al. 2008) and exert antibacterial activity (Martino et al. 1999).

The composition of the globe artichoke phenolic fraction includes four mono-caffeoylquinic isomers, six dicaffeoylquinic isomers, six flavonoid glycosides, and at least seven anthocyanins (Lattanzio et al. 2009). The genes involved in the biosynthesis of the mono-caffeoylquinic acid (chlorogenic acid) have been identified as well as their regulation under UV-C stress (Comino et al. 2007, 2009; De Paolis et al. 2008; Moglia et al. 2009; Menin et al. 2010). Conversely, the biosynthetic pathway leading to di-caffeoylquinic acids is a matter of debate (Villegas and Kojima 1986; Hoffmann et al. 2003; Niggeweg et al. 2004).

The characteristic bitterness of both globe artichoke and cultivated cardoon is mainly due to the presence of sesquiterpene lactones (STLs), of which the two major representatives are cynaropicrin and, at lower concentration, grosheimin and its derivatives (Schneider and Thiele 1974; Cravotto et al. 2005). Cynaropicrin, like many sesquiterpenes lactones, has various medicinal activities (Shimoda et al. 2003; Cho et al. 2004; Schinor et al. 2004; Emendorfer et al. 2005; Ishida et al. 2010) among which cytotoxicity against several types of cancer cells (Yasukawa et al. 2010). In globe artichoke a germacrene A synthase, involved in the first step

of STLs biosynthesis, has been recently isolated, functionally characterized and mapped (Menin et al. 2012).

C. cardunculus has great potential as a source of renewable energy, thanks to its productivity of lignocellulosic biomass. The caloric value of the three *C. cardunculus* taxa is analogous, however cultivated cardoon has the highest biomass yield, which can reach up to ~ 19 t/ha dry matter with an energy value ~ 17 MJ/kg (Angelini et al. 2009; Ierna and Mauromicale 2010; Portis et al. 2010; Ierna et al. 2012). The species has been also identified as a candidate for the production of seed oil which is suitable for both comestible and bio-fuel end-uses. Seed yield in cardoon is about 2 t/ha and up to 0.8 t/ha in globe artichoke (at 5 % w/v moisture), from about 25–30 % of which is oil of good alimentary quality (Foti et al. 1999) due to its high and well balanced content of oleic and linoleic acids, its low content of free acids, peroxides, saturated and linoleic acids and a favourable α -tocopherol content (Maccarone et al. 1999), while the seed material left after oil extraction can be used as a component of animal feed.

22.3 Linkage Analyses: State of the Art

The genome organization of *C. cardunculus* ($2n = 2 \times = 34$; haploid genome size ~ 1.08 Gbp), unlike other species belonging to the Asteraceae family (e.g. sunflower, lettuce and chicory), remains largely unexplored. The species is an out-breeder, and is characteristically highly heterozygous. Its marked level of inbreeding depression inhibits the use of backcross, F_2 or recombinant inbred populations for mapping purposes. As haploid induction—via either androgenesis or gynogenesis—has not yet been achieved (Motzo and Deidda 1993; Chatelet et al. 2005; Stamigna et al. 2005), no possibility is presently available to generate doubled haploid populations. Thus, genetic mapping in the species has relied on a double pseudo-testcross approach (Grattapaglia and Sederoff 1994), in which segregating F_1 progeny are derived from a cross between two heterozygous individuals.

The first genetic maps of *C. cardunculus* were provided by Lanteri et al. (2006), and based on a cross between two genotypes of globe artichoke, namely the varietal types ‘Romanesco C3’ (a late-maturing non-spiny type used as female) and ‘Spinoso di Palermo’ (an early-maturing spiny type used as male). This population was genotyped using a number of PCR-based marker platforms, resulting in a ~ 1300 cM female map consisting of 204 loci, divided into 18 linkage groups (LGs), and a ~ 1200 cM male map comprising 180 loci and 17 LGs. The two maps shared 78 loci, which allowed for the alignment of 16 of the LGs. The maps have since been extended by the inclusion of three genes involved in the synthesis of caffeoylquinic acid (Comino et al. 2009; Moglia et al. 2009) and a number of microsatellite loci, of which 19 were represented in both maps (Acquadro et al. 2009).

New maps have lately been generated from a set of F_1 progeny involving the cross between the same female parent as previously (‘Romanesco C3’) and the cultivated cardoon genotype ‘Altilis 41’ (Portis et al. 2009a). The cultivated cardoon map

comprised 177 loci, subdivided into 17 LGs and spanning just over 1000 cM, while the globe artichoke one featured 326 loci arranged into 20 LGs, spanning ~ 1500 cM with a mean inter-marker distance of ~ 4.5 cM. A set of 84 loci shared between this 'Romanesco C3' map and the previously developed one (Lanteri et al. 2006) allowed for map alignment and the definition of 17 homologous LGs, corresponding to the haploid chromosome number of the species. Later on, the maps have been integrated with the inclusion of all the genes involved in the synthesis of caffeoylquinic acids known in the species (Menin et al. 2010).

Since more markers were needed to saturate the maps, a further wide set of SSR markers was developed from ESTs (expressed sequence tags) of globe artichoke, made available by the Composite Genome Project (CGP; <http://compgenomics.ucdavis.edu/>). Using a custom bioinformatic pipeline, 36,321 ESTs were assembled into 19,055 unigenes (6,621 contigs and 12,434 singletons), annotated, and mined for perfect SSRs. Over 4,000 potential EST-SSR loci, lying within some 3,300 genes (one SSR per 3.6 kbp) have been identified, and PCR primers for the amplification of more than 2,000 of these have been designed. In a test of a sample of 300 of these assays, over half proved to be informative between the parents of the available mapping populations (Scaglione et al. 2009). As a result, a large number of these EST-SSR loci have been integrated into the globe artichoke and cultivated cardoon maps (Portis et al. 2012). The integration of 139 EST-SSR loci has significantly improved the resolution and accuracy of the 'Romanesco C3' and 'Altilis 41' maps. The female map was built with 473 loci spanning 1,544 cM with a mean inter-marker distance of 3.4 cM, corresponding to a 3.8 % increase in length over the earlier map, but in a ~ 28 % decrease in the mean inter-marker distance. The male map consisted of 273 loci spanning 1486 cM, with a mean inter-marker distance of 5.4 cM, representing a marked increase in both length (+ 42 %) and number of loci (+ 50 %), together with a minor decrease in the mean inter-marker distance (-5 %). The two maps shared 66 codominant loci (64 SSRs and two SNPs), which allowed for the alignment of all the 'Romanesco C3' with the 'Altilis 41' LGs. Following alignment a consensus linkage map based exclusively on microsatellite and SNP markers (as depicted in Fig. 22.1) was constructed (Portis et al. 2009b). The consensus map is shown in Fig. 22.2; it comprised 227 loci (217 SSRs and ten SNPs targeting genes involved in the synthesis of caffeoylquinic acids) arranged into 20 LGs (LOD threshold > 6.0). The consensus map length was 1068.0 cM, with a mean inter-marker spacing of 5.2 cM. The length of LGs varied from 4.0–113.7 cM (mean 62.8 cM), with the largest LG containing 36 loci. Lowering the LOD threshold to 5.0 resulted in the merging of three pairs of LG, thereby reducing the overall number to 17, corresponding to the haploid chromosome complement of the species. The majority of the LGs contained a mixture of 'Romanesco C3', 'Altilis 41' and shared co-dominant markers, with only four (LG_9, 13, 14 and 7) carrying shared loci and markers only present in the 'Romanesco C3' map.

Putative functions have been deduced for SSR markers derived from ESTs using homology searches with public protein databases. Annotation of mapped loci was performed via BlastX search as well as InterPro scan and GO categorisation made

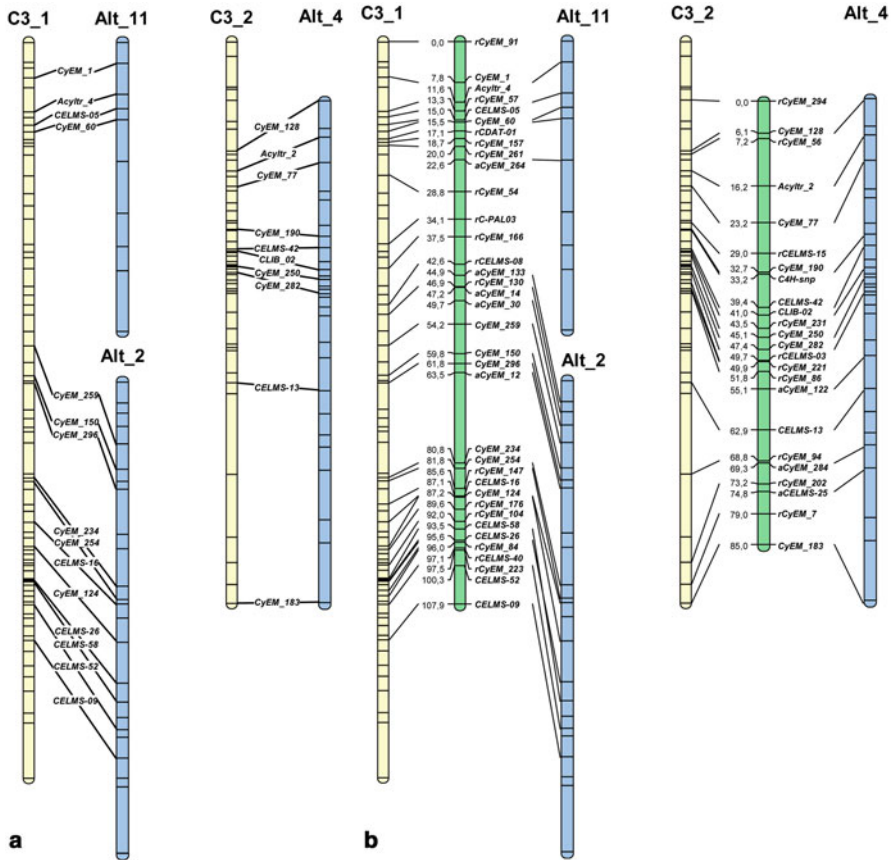


Fig. 22.1 Examples of alignment and consensus LG construction. Alignment of the ‘Romanesco C3’ (yellow) and the ‘Atilis 41’ (blue) LGs based on common markers **a** SSR-based consensus LGs (green) construction **b** ‘r-’ and ‘a-’ indicate markers segregating only in, respectively, ‘Romanesco C3’ and ‘Atilis 41’. Marker nomenclature is the one reported in Portis et al. (2009a) and Scaglione et al. (2009)

it possible to tag some biological functions. A set of 17 EST-SSR markers were annotated with GO terms involved in the ‘response to stimulus’ (Table 22.1), five and eight of which were derived from transcripts related to response to cold and salt stress, respectively. As an example, the marker CyEM-42, developed from the contig CL4773Contig1 (Scaglione et al. 2009) and mapped on LG_12 of the SSR-based consensus map, showed high aminoacidic similarity (81 %) with the protein kinase PBS1 of *Arabidopsis*. To consider reliable orthology, a reciprocal tblastx was performed, and no better alignment than that of contig CL4773 was detected. PBS1 was found to work as R gene against the bacterial pathogen *Pseudomonas syringae*, where its cleavage, operated by the pathogens’ effector AvrPphB, triggers

Table 22.1 CyEM (Cynara Expressed Microsatellites) markers with Gene Ontology annotation for stimuli response-related terms

GO ID	Term	N of loci	EST-SSR loci
GO:0050896	response to stimulus	17	CyEM -008, -030, -42, -054, -057, -070, -072, -093, -120, -135, -145, -150, -152, -218, -229, -259, -266
GO:0009628	response to abiotic stress	13	CyEM -008, -030, -054, -070, -093, -120, -135, -145, -150, -152, -218, -229, -259
GO:0042221	response to chemical stimulus	4	CyEM -093, -218, -229, -266
GO:0006950	response to abiotic stress	15	CyEM -008, -030, -054, -057, -070, -072, -093, -120, -135, -145, -150, -152, -229, -259, -266
GO:0009266	response to temperature stress	5	CyEM -008, -054, -093, -145, -150
GO:0006970	response to osmotic stress	8	CyEM -030, -070, -093, -120, -135, -152, -229, -259
GO:0010033	response to organic substance	3	CyEM -093, -229, -266
GO:0009409	response to cold stress	5	CyEM -008, -054, -093, -145, -150
GO:0009651	response to salt stress	8	CyEM -030, -070, -093, -120, -135, -152, -229, -259

the signalling cascade, generating the host response (Shao et al. 2002). *Pseudomonas* spp. together with other endophytic bacteria may affect globe artichoke plants both in field and during micropropagation (Penalver et al. 1994) and the CyEM-42 may be likely considered a reliable marker for tagging a bacterial resistance trait in the species. On the whole, these EST-SSR markers may be defined as functional markers with the potential to target polymorphisms in gene responsible for traits of interest; they can be also particularly useful for constructing comparative framework maps with other Asteraceae, giving the possibility to amplify ortholog genes and provide anchor loci.

This SSR-based consensus map of *C. cardunculus* is based on a robust marker platform of SSRs and a few gene-based SNP loci. It is expected that the further positioning of markers within target regions will provide key tools for marker-assisted breeding programs as well as the necessary framework to exploit mapping data obtained from diverse populations. At present, ~200 of the loci on the consensus map (about 88 %) are sited within genic sequence, presenting some opportunity to identify candidate genes for particular traits within the species.

22.4 SNP Mining

The first set of SNP (single nucleotide polymorphism) markers available for the species has been developed on genes involved in the synthesis of caffeoylquinic acid, as above reported. The allelic forms of globe artichoke acyltransferases HCT,

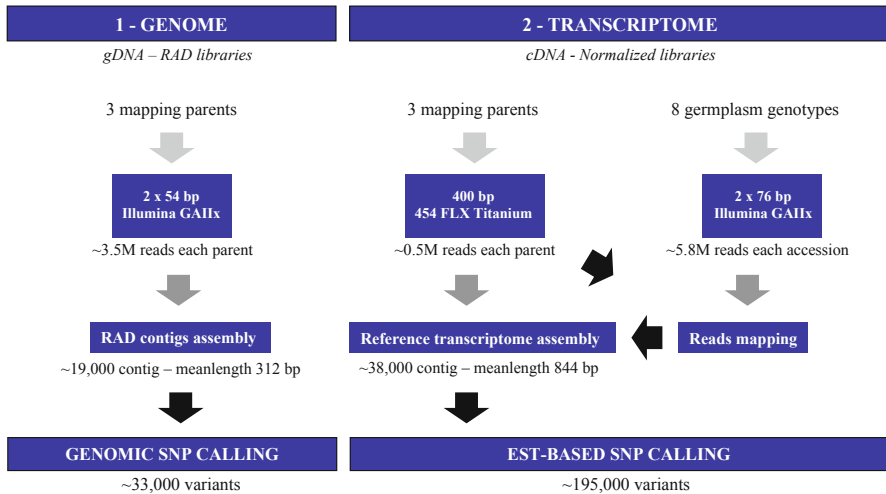


Fig. 22.3 SNP mining workflow in *Cynara cardunculus*

HQT (Comino et al. 2007, 2009) and the hydroxylase C3'H (Moglia et al. 2009), were analysed in the two globe artichoke parental genotypes ('Romanesco C3' and 'Spinoso di Palermo') of the first genetic maps (Lanteri et al. 2006) and SNPs were identified. SNP genotyping of the F₁ progeny was carried out with the tetra-primers ARMS-PCR method (Ye et al. 2001; Chiapparino et al. 2004). A further SNP set has been later on developed by Menin et al. (2010) on three acyltransferases and on the *C4H*, *4CL* and *MYB12* genes, identified by an *in silico* scan of the globe artichoke unigene set assembled by Scaglione et al. (2009). Gene homologues were re-sequenced in the parental genotypes (globe artichoke 'Romanesco C3' and cultivated cardoon 'Atilis 41') of the genetic maps developed by Portis et al. (2009a) and genes successfully mapped.

Recent advances in next-generation DNA sequencing technologies have made possible the development of high-throughput SNP genotyping platforms, that allow for the simultaneous interrogation of thousands of SNPs. Such resources have the potential to facilitate the rapid development of high-density genetic maps, and to enable genome-wide association studies as well as molecular breeding approaches in a variety of taxa (Bachlava et al. 2012). Thousands of SNPs have been recently developed in *C. cardunculus* by Next-Generation Sequencing (NGS) technology using two complementary approaches (Fig. 22.3):

1. genomic RAD (Restriction-site Associated DNA) tag sequencing (Miller et al. 2007) in combination with the Illumina Genome Analyzer sequencing device (Baird et al. 2008) of three genotypes (globe artichoke, cultivated cardoon and wild cardoon) that were crossed for developing F₁ mapping populations (Scaglione et al. 2012a);

- transcriptome sequencing, via 454 and Illumina technologies, of the same three genotypes plus eight, five of which were globe artichoke, two cultivated and one wild cardoon (Scaglione et al. 2012b). Alongside, a functional characterisation and annotation of the obtained sequence set was performed. These SNPs represent a one-stop resource to produce a dense *C. cardunculus* genetic map via high-throughput genotyping technologies.

22.4.1 Genomic SNP Mining

The recently developed restriction-site associated DNA (RAD) approach (Box 1) has been combined with the Illumina DNA sequencing platform to enable the rapid and mass discovery of SNP markers. Three genomic RAD libraries were obtained from the three *C. cardunculus* genotypes belonging to the three taxa of the species and parents of two mapping populations. The first mapping population is the F₁ progeny involving the cross between globe artichoke ('Romanesco C3', female parent) and cultivated cardoon (genotype 'Atilis 41') (Portis et al. 2009a). The second one is the F₁ progeny involving the cross between the same female parent as previously and the wild cardoon (genotype 'Creta 4') (Lanteri et al. 2011).

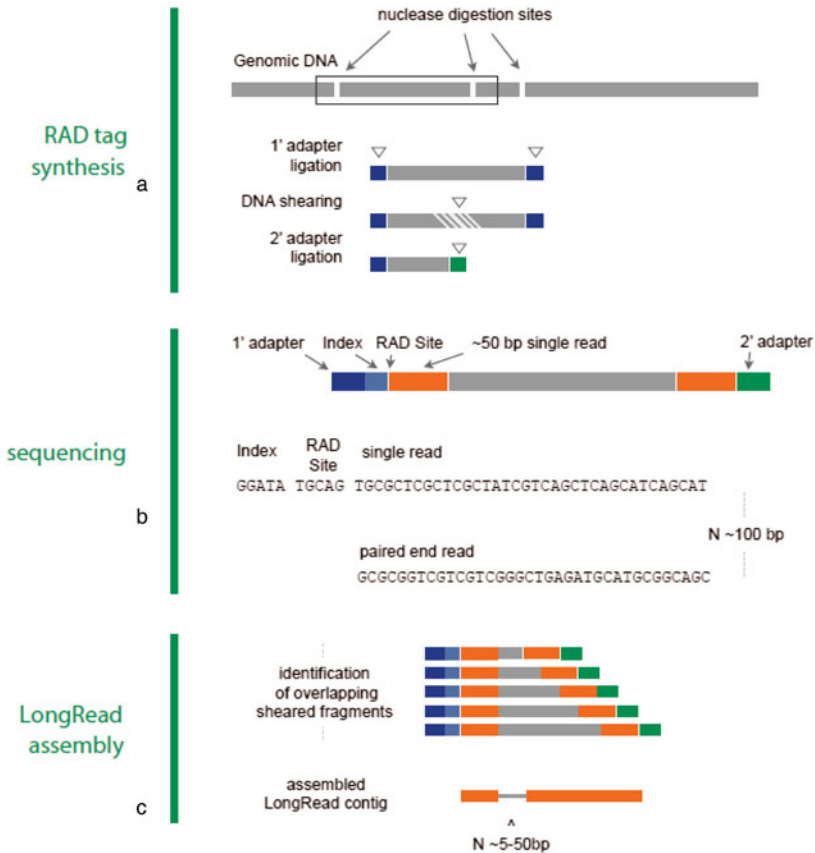
22.4.1.1 RADseq (Restriction-site Associated DNA sequencing)

An efficient approach for SNP discovery is RAD "Restriction-site Associated DNA" (Miller et al. 2007), coupled with NGS technologies (Baird et al. 2008), which has been recently termed as RADseq (Davey et al. 2011). At least 65 papers have been recently published in both animals (snails, moths and salmon, sturgeon, butterflies, beetles and worms) and plants (ryegrass, oaks, lolium, eggplant and globe artichoke). A detail review is available at the wiki RAD-sequencing page (University of Edinburgh; <https://www.wiki.ed.ac.uk/display/RADSequencing>).

The strategy requires the enzymatic digestion of a genome with at least one restriction enzyme and the sequencing of the resulting fragments through an Illumina Genome Analyzer. The fragments from one sample are ligated to a modified Illumina adapter containing a unique identifying sequence (Molecular Identifier, or MID). A list of the available primers can be found at the above-cited wiki RAD-sequencing section. The fragments from many samples (e.g. a mapping population) can consequently be pooled together and sequenced on a single lane. The resulting reads can be segregated using the MID present at the start of each read. By sequencing simultaneously all the individuals of a population of interest, and by comparing the tags, thousand of SNPs at different genetic loci can be identified in a single experiment.

The protocol is depicted in the figure reported below. (A) Genomic DNA is digested with a restriction enzyme and a barcoded P1 adapter is ligated to the fragments. The P1 adapter contains a forward amplification primer site, an Illumina sequencing primer site, and a barcode for sample identification. Adapter-ligated fragments are

pooled (if multiplexing), sheared, size-selected (e.g. 300–800 bp) and ligated to a second adapter (P2). The P2 adapter is a divergent “Y” adapter, containing the reverse complement of the P2 reverse-amplification primer site, preventing amplification of genomic fragments lacking a P1 adapter. (B) The samples are analysed on an Illumina Genome Analyzer IIx following the paired ends (2×54 bp, or more) genomic DNA sequencing protocol. The generated sequences are then sorted according to their multiplex identifier tag (barcode). (C) The sequences are de novo assembled using a bioinformatics DNA assembler (e.g.: Velvet). Assembled LongRead® contigs can be generated by a set of algorithms developed at Floragenex Inc. (Oregon, USA).



22.4.1.2 RAD tag Sequencing and de novo Assembly

The RAD-seq exercise produced 9.7 million reads (19.4 million Pair End—PE), equivalent to ~ 1 Gbp of sequence. The distribution of reads was uneven across the three DNA samples, with 1.2 million reads achieved for globe artichoke, 2.6 million for cultivated cardoon and 5.9 million for wild cardoon; the latter, being the largest

set, was chosen as the basis for *de novo* contigs assembly. The assembly procedure created 19,061 reference genomic contigs, spanning 6.11 Mbp (with N50 = 321 bp and a mean a contig length of 312 bp). The GC content was ~ 37.4 % which is similar to that of many dicots species (Jaillon et al. 2007) and represents the first survey on the base composition of the *C. cardunculus* genome.

22.4.1.3 RAD tag Annotation

The contig sequences characterisation was conducted using the BlastX algorithm and it resulted in the annotation of 5,335 contigs (28.0 %). Regardless of the genome-wide RAD sampling, a noteworthy part of the annotated sequences might be represented by coding regions, since a methylation-sensitive enzyme (*Pst*I) was used to produce the RAD-tag libraries (Palmer et al. 2003), although the rather short length of the RAD contigs made difficult to distinguish between putative genes and pseudogenes. Enzyme codes were retrieved for 1,327 contigs, defining a unique set of 313 putative enzymatic activities, which were mapped onto KEGG reference pathways (<http://www.genome.jp/kegg/>). The remaining portion of the contig set (72 %) was not attributed to any known sequence, likely due to the RAD contigs shortness.

The transposable DNA element footprints detected, using RepeatMasker software (v3.2.9; <http://www.repeatmasker.org>) implemented with the RMBlast algorithm, and adopting the *Viridiplantae* repeats as reference, accounted for a 0.2 % of the sequence, while 1.2 % of the sequences derived from LTR retroelements, including Ty/Copia-like (0.8 %) and Gypsy-like (0.2 %). This quantification of transposable element abundance could have been underestimated, but these data represents a useful snapshot of relative abundance of each different mobile element class in *C. cardunculus*.

22.4.1.4 SNP Calling

The PE sequences generated for each mapping parent were aligned using the reference contig set as a scaffold. In total, ~ 33,000 sequence variants were detected, including 1,520 short indels, distributed over 12,068 contigs. The overall SNP frequency was estimated to be 5.6 per 1,000 nucleotides, a level which is almost equal to that found in the non-coding regions of the *V. vinifera* genome (5.5 per 1,000 nucleotides; Velasco et al. 2007) and very similar to that observed in *Citrus* spp. ESTs (6.1 per 1,000 nucleotides; Jiang et al. 2010). A subset of ~ 17,400 SNPs was obtained considering allelic variant which were informative for both mapping populations (16,727 SNPs, and 723 1–2nt indels) distributed over 7,478 contigs.

Since *C. cardunculus* is highly heterozygous, SNPs were categorized as intra- or inter-varietal, where the former also represents the heterozygous state of the analysed genotype. The two types were not exclusive, therefore heterozygous SNPs present in one sample could be found in both heterozygous or homozygous states in other genotypes. The number of heterozygous SNP loci was 1,235 in the globe artichoke,

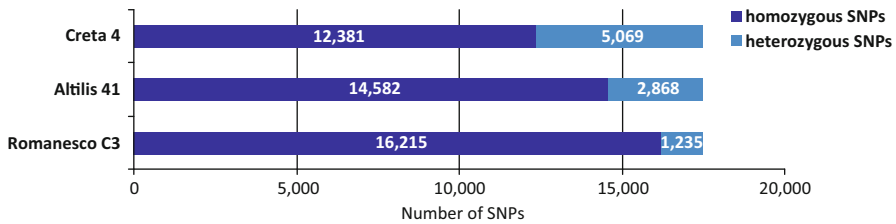


Fig. 22.4 Proportion of heterozygous SNPs across the three mapping parents

Table 22.2 454-derived sequencing and assembly. The output statistics were calculated following the removal of contaminating and adaptor sequences. Data are intended after quality filtering and sequence clipping

#	Genotype	<i>C. cardunculus</i> taxon	Sequencing results			Assembly results	
			Raw reads (M)	Total (Mbp)	Mean length (bp)	Contigs	Mean length/N50 (bp)
1	'Romanesco C3'	var. <i>scolymus</i>	0.43	184	421	37,622	834/723.8
2	'Altilis 41'	var. <i>altilis</i>	0.61	246	402	40,130	761/699.9
3	'Creta 4'	var. <i>sylvestris</i>	0.69	263	377	42,837	772/688.5
<i>Total</i>			1.74	693	392	38,726 ^a	951/844.3 ^a

^aAsterisks indicate results obtained by merging the three independent assemblies (see Fig. 22.3)

2,868 in the cultivated cardoon and 5,069 in the wild cardoon mapping parents (Fig. 22.4). Heterozygous SNPs are of key importance for mapping studies since for the linkage analysis a two-way pseudo-testcross approach, based on a segregant F₁ progeny, was adopted. In this sense, a key parameter for the successful isolation of such useful SNP markers was the sequencing coverage.

22.4.2 Transcriptomic SNP Mining

A total of eleven *C. cardunculus* EST libraries were produced and after the normalisation procedures, they were separately sequenced. Three libraries, deriving from the three mapping parents (Table 22.2), were sequenced with the 454 Titanium (Roche) to produce a reference transcriptome. Eight libraries, set up from five globe artichoke accessions, two cultivated cardoon and one wild cardoon genotypes (Table 22.3), were sequenced using the Illumina GAIIx platform, in order to highly increase the total SNP calling amount.

22.4.2.1 EST Sequencing and de novo Assembly

The outcome of 454-based cDNA sequencing of the three mapping parents generated some 1.7 M reads of overall length 695 Mb, which were reduced to 692 Mb after a post-sequencing filtering. The mean read length was equal to 392 bp (Table 22.2).

Table 22.3 GAIIX (Illumina)-derived sequencing. A total of 46.5 M raw reads were generated in two GAIIX lanes and 6.7 Gbp were retained after removing adaptor and contaminating sequences. The windowed quality clipping routine produced a final dataset of 6.2 Gbp. A higher number of bases were obtained for single ends, because 84 sequencing cycles were used instead of the 76 used for the paired ends

#	Genotype	<i>C. cardunculus</i> taxon	Raw reads (M)	First mates (Mbp)	Paired mates (Mbp)
4	'Romanesco Zorzi'	var. <i>scolymus</i>	6.6	458	408
5	'Violetto di Chioggia'	var. <i>scolymus</i>	6.6	470	420
6	'Violetto Pugliese'	var. <i>scolymus</i>	5.2	367	331
7	'Spinoso Sardo'	var. <i>scolymus</i>	6.7	474	424
8	'Imperial Star'	var. <i>scolymus</i>	6.4	459	415
9	'Blanco de Peralta'	var. <i>altilis</i>	4.8	340	305
10	'Gobbo di Nizza'	var. <i>altilis</i>	5.6	380	341
11	'Sylvestris_LOT23'	var. <i>sylvestris</i>	4.6	322	287
<i>Total</i>			46.5	3,271	2,931

cDNA libraries of other eight genotypes (Table 22.3) were sequenced using a GAIIX platform (Illumina) producing 6.9 Gbp of raw data (46.4 M paired-end reads) with a mean of 5.8 M reads per accession. The data set was reduced to 6.7 Gbp following the removal of adaptor sequences and other contaminants, and it was further reduced to 6.2 Gbp after quality trimming. For the *de novo* assembly process only the 454 reads were considered, while the Illumina data were simply adopted to increase the efficiency of the SNP mining process.

The assembly of 454 reads was achieved by a two-tier approach using the MIRA assembler ver.3.2.0 (Chevreux et al. 2004). In a first step, each individual sample was assembled independently. The process generated 37,622 contigs for 'Romanesco C3', 40,130 contigs for 'Altilis 41', and 42,837 contigs for 'Creta 4' with N50 contig lengths of 834 bp, 761 bp and 772 bp, and mean coverage levels of 7.31, 8.45 and 9.17X, respectively. For the 'Romanesco C3' assembly, a subset of 11,276 contigs resulted from the incorporation of a prior set of 28,641 Sanger ESTs (www.ncbi.nlm.nih.gov/dbEST). Then, after contaminant removal by BLASTX analysis, the three datasets were merged into a set of 38,726 contigs. This "reference" assembly spanned 32.7 Mbp and had a GC content of 42.1 %. The mean contig length was 844.3 bp (N50: 951 bp).

A second assembly phase was carried out by merging at least two *taxon*-derived contigs from the first phase, and 20,469 contigs were generated. They consisted of a subset with a mean length of 1054 bp, while 5,375, 6,669 and 6,213 remained as single *taxon*-derived contigs of var. *scolymus*, var. *altilis* and var. *sylvestris*, respectively.

22.4.2.2 Sequence Analysis and Functional Annotation

The sequence reads were assembled into 38,726 reference transcripts, which were successfully annotated, using the Blast2GO pipeline, by gene ontology terms via Blast and InterPro analyses. Enzymes were tagged on KEGG's reference pathways (www.genome.jp/kegg/), including primary and secondary metabolisms. On

the whole, 16,419 enzyme codes were retrieved (12,449 transcripts) and subsequently mapped onto KEGG's pathways. The sample of *C. cardunculus* enzymes consisted of 1,133 unique enzyme codes distributed across 147 pathways. To provide an example, by analyzing the whole transcriptome complement, a subset of 71 enzymatic activities involved in phenylpropanoid synthesis were identified; 21 of these were annotated at varying levels of redundancy in the core phenylpropanoid pathway (KEGG's map: 00940), in which the synthesis of caffeoylquinic and di-caffeoylquinic acids (CQAs and dCQAs) takes place (Fig. 22.5).

Transcriptional factor function was assigned to 1,398 transcripts, scattered across 67 families, while 316 sequences were tagged as candidate Resistance Gene Analogs (RGAs). Each sequence was scanned for the presence of recognition sites for known plant miRNAs. In total, target annealing sites for 302 miRNAs were located in 1,043 transcripts, which mainly belong to the categories: "defense response" and "programmed cell death/apoptosis", "reproduction", "development of anatomical structure", "photosynthesis", "transmembrane receptor activity" and "transcription factor activity". The 454-based assembly included non-nuclear transcripts. The *C. cardunculus* chloroplast genes identification was based on similarity to those of lettuce and sunflower (Timme et al. 2007) leading to the categorization of 137 contigs, of which 80 were putatively assigned the chloroplast genome. Similarly, the grapevine (*Vitis vinifera*) mitochondrial genes (Goremykin et al. 2009) aided in the identification of 52 *C. cardunculus* contigs, which were putatively attributed to the mitochondrial genome.

To estimate the transcriptome representation and its gene-level redundancy (e.g. splicing variants), two different approaches were adopted. Using the *A. thaliana* gene content, the 454 sequencing output was predicted to be assembled in a total of 29.3 Mbp, distributed in 24,064 unigenes (average length of 1,216 bp) which covered 96 % of the transcriptome. Alternatively, the final contig set (38,000) was clustered by collapsing gene variants (e.g. alternative splicing), which generated a set of 29,830 unigenes that represents a *bona fide* estimation of the gene content of *C. cardunculus*. Data suggest that 23 % of splicing variants could be present in the transcriptome assembly.

22.4.2.3 Read Mapping and SNP Calling

About 1.5 M of the 454-derived reads were aligned to the reference contig set (38,726 contigs). This number was reduced to ~ 1.0 M by removing those that showed more than one unique alignment, thereby lowering the risk of false SNP calls due to misalignment of paralog-derived reads or to redundancy resulting from splicing variants. The same procedure was repeated for the Illumina-derived reads, producing an alignment of ~ 60 M paired ends. Resolving paired ends reduced this to a set of ~ 21 M reads.

An assembly based on about 35 M sequences was generated by merging the 454 and Illumina sequence datasets, resulting in a median reference transcriptome coverage of 96X with 26,990 reference contigs containing at least 20 mapped reads.

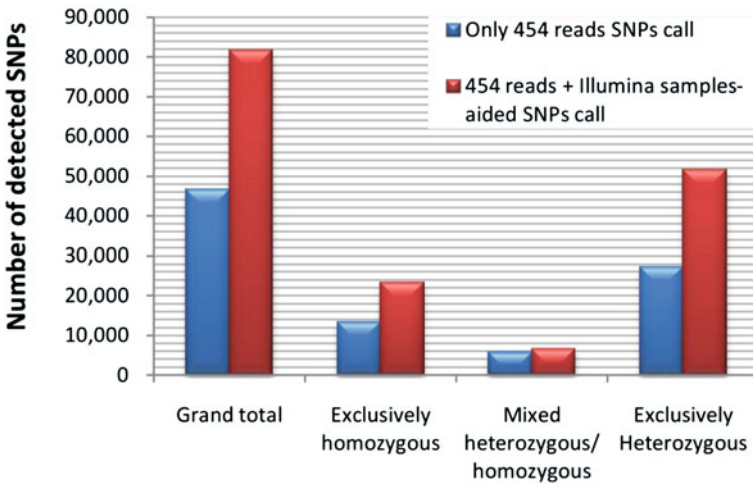


Fig. 22.6 Combined calling of SNPs. The number of calls based solely on the 454-derived reads is shown in *blue*, and the combined SNP discovery based on both the 454- and the Illumina-based sequence in *red*. “*Exclusively homozygous*” and “*exclusively heterozygous*” refer to allelic variants present in only one of the three 454-sequenced libraries

Reliable SNPs (Bayesian probability > 95 %) were detected at 195,400 sites across the set of 11 accessions. The average SNP frequency was calculated at one per 167 bp, with a mean of five per contig. Each SNP site was interrogated by scoring for the presence of at least one accession-specific sequence. Sequence information was available from an average of nine accessions per SNP site, and a core subset of 57,125 SNPs showed coverage from all the samples. The merging of the Illumina-derived reads (eight accessions) with 454-generated reads substantially increased the number of parent-specific SNPs that were identified (Fig. 22.6).

SNP frequency in the *C. cardunculus* transcriptome appears to be comparable to that found in the heterozygous grapevine whole genome sequence (Velasco et al. 2007) and among *Citrus* ssp. ESTs (Jiang et al. 2010). Overall, SNPs were most frequent in 3'-UTR (one per 126 bp), followed by the CDS (one per 169 bp), and the 5'-UTR (one per 265 bp). Within the UTRs, the frequency also matched that obtained in tomato expressed sequence (Jimenez-Gomez and Maloof 2009), while it was markedly different to that present in the coding region (~2 per kb). This discrepancy may reflect either the greater tolerance by the heterozygous state of non-synonymous substitutions, or merely is an ascertainment bias due to the analysis of a larger germplasm panel which also included accessions of a wild relative.

In *C. cardunculus*, as previously pointed out, the presence of intra-accession allelic variation is of particular interest. As expected by their shallower coverage, the 454-derived sequences produced a somewhat lower frequency of SNPs with successful heterozygous SNP calling (Fig. 22.7). ‘Altilis 41’ was relatively the least heterozygous of the accessions (17,570 loci), as has been observed previously (Portis et al. 2005b, 2009a), while ‘Romanesco Zorzi’ was the most heterozygous (43,387 loci),

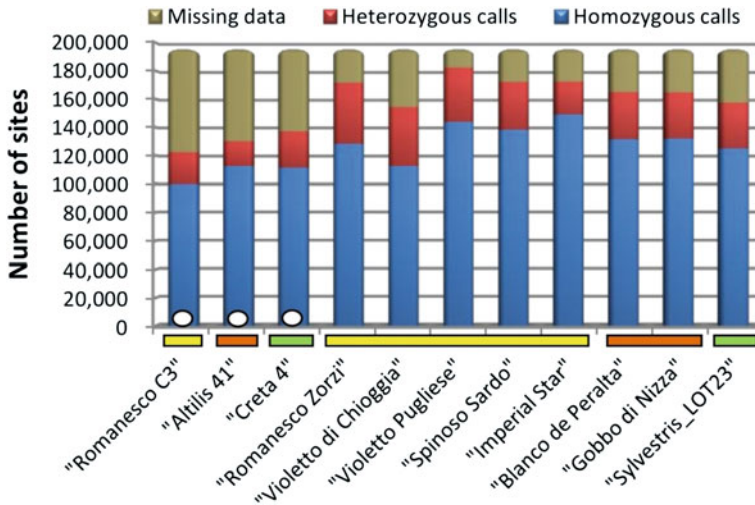


Fig. 22.7 The allelic state at SNP loci. Bars indicate the total number of SNP loci in the homozygous or heterozygous state (or missing) for each accession. Each bar's colour identifies the *C. cardunculus* taxa (green = sylvestris, orange = altilis, yellow = scolymus). White dots identify the three accessions sequenced using 454 technology

followed by 'Violetto di Chioggia' (41,824 loci). 'Imperial Star' had the lowest ratio of heterozygous variants among globe artichoke genotypes (13.5%), which likely reflects its development from crosses among less genetically differentiated genotypes.

22.4.3 Conclusions

The second generation technologies provide high sequencing throughputs at significantly reduced costs if compared to Sanger. These platforms are currently employed for large-scale SNP discovery projects and, for medium-scale projects, they have been frequently applied in combination with reduced-complexity libraries, targeting genomic subsets.

One such method, aimed at decreasing the sample complexity, is to build up a genomic library, with a reduced locus representation including only a subset of sequences generated by restriction enzymes, which cut at frequent intervals throughout the genome. The generation of a SNP set can be achieved through the deep-sequencing of such libraries and the comparison between allelic variants can identify thousands of SNPs.

The recent RAD (Baird et al. 2008) approach is focused on the targeting of a discrete number of genomic regions adjacent to specific restriction sites, and it can effectively reduce the number of the fragments to be sequenced in a given complex

genome. This strategy (see 22.4.1.1) represents a promising experimental scheme in term of costs and technical simplicity and, so far, has been successfully adopted for SNP discovery in many plant and animal species (Davey et al. 2011).

An alternative approach is to focus onto the transcriptome deep-sequencing, which reduces the representation of low information-content repetitive sequences in species possessing a large genome and/or without a finished genome sequencing project. An EST library can lead to identify a large number of genetic loci and targeting SNPs in coding sequences. This kind of library represents a one-stop resource useful for many downstream applications and to address many biological questions in plant science. It can aid the identification of genes underlying phenotypes of interest through the development of expression arrays or provide thousands of loci as a source of potential markers for QTL mapping applications and population genomic studies.

The two experimental workflows led to produce a massive set of SNPs in *C. cardunculus*, and made possible to create an extensive gene catalogue, as a valuable resource for upcoming genomic and genetic studies. Both approaches have proven to be efficient for SNP mining, although characterized by peculiarities and limitations which deserve to be considered in view of specific research targets. In *C. cardunculus* the EST sequencing approach generated a set of reference coding sequences spanning 32.7 Mbps, establishing a 'general gene catalog' of 38,726 as *bona fide* representation of the transcriptome. In contrast the RAD-tag sequencing approach permitted to sequence 6.0 Mbps separated in lesser and shorter number of contigs (~ 19,000; 28 % of which were annotated as CDS-like). The number of SNPs was higher for EST than for RAD-tag approach (195,000 vs. 34,000); nevertheless, the SNP frequency observed in the two pipelines were somewhat comparable (5.9 vs. 5.6 per 1,000 nt). The RAD-tags data revealed to be extremely informative to preliminary survey the repetitive DNA component of the *C. cardunculus* genome, and allowed us to make some inferences regarding the contribution of DNA methylation in inhibiting its expansion (Scaglione et al. 2012a).

From the standpoint of costs, RAD technology was attempted with a great technical simplicity and a low cost/time expense. The cDNA library setting up was indeed more complex for both the need of standardization/normalisation procedures and some extra enzymatic steps required, however, side by side, its sequencing output provided a better picture of the globe artichoke coding genome. Bearing in mind a future in which the globe artichoke genome will be completely sequenced and publicly available, the genomic RAD approach may represent one of the most feasible and cheap strategy for accomplishing affordable targeted re-sequencing projects. It is also likely that the increasingly lowering of sequencing costs will make the scientific community to converge towards new approaches of 'genotyping-by-sequencing'. This scheme proceeds to explore all the nucleotidic positions of a genome in a single experiment, and will permit an integration of mapping and sequencing steps, likely bypassing many costly physical mapping procedures.

The combination of two NGS platforms (454 FLX Titanium—Roche and GAIIX—Illumina) for the extensive characterization of the genome and transcriptome of *C. cardunculus*, has proven to be a highly reliable tools for SNP discovery. Overall, the availability of such a large number of sequence-based markers, in a format allowing for high-throughput genotyping, offers opportunities to developed a high-density

genetic map and association mapping studies aimed at correlating molecular polymorphisms with variation in phenotypic traits, as well as for molecular breeding approaches in a species which has multiple end-uses such as food, nutraceuticals and bioenergy. The high number of mined SNPs represents also an excellent resource for evolutionary genetic studies in cultivated forms and their wild relative as well as for comparative genetic mapping studies aimed at understanding patterns of genome rearrangement between *C. cardunculus* and related species.

Acknowledgements

We wish to thank:

- Loren H. Rieseberg, Steven J. Knapp and Zhao Lai (Compositae Genome Project) for founding the RAD tag and transcriptome sequencing within the U.S. National Science Foundation grants “Comparative genomics of phenotypic variation in the Compositae (DBI-0820451)”
- Giovanni Mauromicale and Rosario P. Mauro (Dipartimento di Scienze delle Produzioni Agrarie e Alimentari (DISPA)—Sez. Scienze Agronomiche, University of Catania) for the development and maintenance in field of the mapping progenies.

References

- Acquadro A, Portis E, Lee D et al (2005) Development and characterization of microsatellite markers in *Cynara cardunculus* L. *Genome* 48:217–225
- Acquadro A, Lanteri S, Scaglione D et al (2009) Genetic mapping and annotation of genomic microsatellites isolated from globe artichoke. *Theor Appl Genet* 118:1573–1587
- Angelini LG, Ceccarini L, Di Nasso NNO, Bonari E (2009) Long-term evaluation of biomass production and quality of two cardoon (*Cynara cardunculus* L.) cultivars for energy use. *Biomass Bioenerg* 33:810–816
- Bachlava E, Taylor CA, Tang S et al (2012) SNP discovery and development of a high-density genotyping array for sunflower. *PLoS one* 7:e29814
- Baird N, Etter P, Atwood T et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS one* 3:e3376
- Barbagallo RN, Chisari M, Spagna G et al (2007) Caseinolytic activity expression in flowers of *Cynara cardunculus* L. *Acta Hort* 730:195–199
- Blumenthal M, Goldberg A, Brinckmann J (eds) (2000) Artichoke leaf. *Herbal medicine: expanded commission E monographs, vol 10. Integrative Medicine Communications, Boston*, pp 210–12
- Basnizki J, Zohary D. (1994) Breeding of seed-planted artichoke. *Plant Breed Rev* 12:253–269
- Brown J, Rice-Evans C (1998) Luteolin-rich artichoke extract protects low density lipoprotein from oxidation in vitro. *Free Rad Res* 29:247–255
- Bundy R, Walker A, Middleton R et al (2008) Artichoke leaf extract (*Cynara scolymus*) reduces plasma cholesterol in otherwise healthy hypercholesterolemic adults: a randomized, double blind placebo controlled trial. *Phytomedicine* 15:668–675
- Chatelet P, Stamigna C, Thomas G (2005) Early development from isolated microspores of *Cynara cardunculus* var. *scolymus* (L.) Fiori. *Acta Hort* 681:375–380
- Chevreux B, Pfisterer T, Drescher B et al (2004) Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14:1147–1159
- Chiapparino E, Lee D, Donini P (2004) Genotyping single nucleotide polymorphisms in barley by tetra-primer ARMS-PCR. *Genome* 47:414–420
- Cho J, Kim A, Jung J et al (2004) Cytotoxic and pro-apoptotic activities of cynaropicrin, a sesquiterpene lactone, on the viability of leukocyte cancer cell lines. *Eur J Pharmacol* 492:85–94

- Comino C, Lanteri S, Portis E et al (2007) Isolation and functional characterization of a cDNA coding a hydroxycinnamoyltransferase involved in phenylpropanoid biosynthesis in *Cynara cardunculus* L. BMC Plant Biol 7:14
- Comino C, Hehn A, Moglia A et al (2009) The isolation and mapping of a novel hydroxycinnamoyltransferase in the globe artichoke chlorogenic acid pathway. BMC Plant Biol 9:30
- Cravotto G, Nano G, Binello A et al (2005) Chemical and biological modification of cynaropicrin and grosheimin: a structure-bitterness relationship study. J Sc Food Agric 85:1757–1764
- Davey JW, Hohenlohe PA, Etter PD et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510
- De Paolis A, Pignone D, Morgese A et al (2008) Characterization and differential expression analysis of artichoke phenylalanine ammonia-lyase-coding sequences. Phys Plant 132:33–43
- Emendorfer F, Emendorfer F, Bellato F et al (2005) Antispasmodic activity of fractions and cynaropocrin from *Cynara scolymus* on guinea-pig ileum. Biol Pharm Bull 28:902–904
- Foti S, Mauromicale G, Raccuia S et al (1999) Possible alternative utilization of *Cynara* spp. I. Biomass, grain yield and chemical composition of grain. Ind Crop Prods 10:219–228
- Gebhardt R (1997) Antioxidative and protective properties of extracts from leaves of the artichoke (*Cynara scolymus* L.) against hydroperoxide-induced oxidative stress in cultured rat hepatocytes. Tox Appl Pharm 144:279–286
- Gebhardt R (1998) Inhibition of cholesterol biosynthesis in primary cultured rat hepatocytes by artichoke (*Cynara scolymus* L.) extracts. J Pharm Exp Therapy 286:1122–1128
- Goremykin V, Salamini F, Velasco R, Viola R (2009) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. Mol Biol Evol 26:99–110
- Grattapaglia D, Sederoff R (1994) Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross—mapping strategy and RAPD markers. Genetics 137:1121–1137
- Hoffmann L, Maury S, Martz F et al (2003) Purification, cloning, and properties of an acyltransferase controlling shikimate and quinate ester intermediates in phenylpropanoid metabolism. J Biol Chem 278:95–103
- Ierna A, Mauromicale G (2010) *Cynara cardunculus* L. genotypes as a crop for energy purposes in a Mediterranean environment. Biomass Bioenergy 34:754–760
- Ierna A, Mauro RP, Mauromicale G (2012) Biomass, grain and energy yield in *Cynara cardunculus* L. as affected by fertilization, genotype and harvest time. Biomass Bioenergy 36:404–410
- Ishida K, Kojima R, Tsuboi M et al (2010) Effects of artichoke leaf extract on acute gastric mucosal injury in rats. Biol Pharm Bull 33:223–229
- Jaillon O, Aury J, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467
- Jiang D, Ye Q, Wang F, Cao L (2010) The mining of *Citrus* EST-SNP and its application in cultivar discrimination. Agricultural Sciences in China 9:79–190
- Jimenez-Gomez J, Maloof J (2009) Sequence diversity in three tomato species: SNPs, markers, and molecular evolution. BMC Plant Biol 9:85
- Kraft K (1997) Artichoke leaf extract—Recent findings reflecting effects on lipid metabolism, liver and gastrointestinal tracts. Phytomed 4:369–378
- Lanteri S, Di Leo I, Ledda L et al (2001) RAPD variation within and among populations of globe artichoke cultivar 'Spinoso sardo'. Plant Breed 120:243–246
- Lanteri S, Saba E, Cadinu M et al (2004) Amplified fragment length polymorphism for genetic diversity assessment in globe artichoke. Theor Appl Genet 108:1534–1544
- Lanteri S, Acquadro A, Comino C et al (2006) A first linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus* L.) based on AFLP, S-SAP, M-AFLP and microsatellite markers. Theor Appl Genet 112:1532–1542
- Lanteri S, Portis E, Acquadro A et al (2011) Morphology and SSR fingerprinting of newly developed *Cynara cardunculus* genotypes exploitable as ornamentals. Euphytica 184:311–321

- Lattanzio V, Kroon PA, Linsalata V, Cardinali A (2009) Globe artichoke: A functional food and source of nutraceutical ingredients. *J Func Foods* 1:131–144
- Maccarone E, Fallico B, Fanella F et al (1999) Possible alternative utilization of *Cynara* spp. II. Chemical characterization of their grain oil. *Ind Crops Prod* 1:229–237
- Martino V, Caffini N, Phillipson J et al (1999) Identification and characterization of antimicrobial components in leaf extracts of globe artichoke (*Cynara scolymus* L.). *Acta Hort* 501:111–114
- Marushia RG, Holt JS (2006) The effects of habitat on dispersal patterns of an invasive thistle, *Cynara cardunculus*. *Biol Invas* 8:577–593
- Matsui T, Ogunwande IA et al (2006) Anti-hyperglycemic potential of natural products. *Mini Rev Med Chem* 6:349–356
- Mauro R, Portis E, Acquadro A et al (2008) Genetic diversity of globe artichoke landraces from Sicilian small-holdings: implications for evolution and domestication of the species. *Cons Genet* 10:431–440
- McDougall B, King P, Wu B et al (1998) Dicafeoylquinic and dicafeoyltartaric acids are selective inhibitors of human immunodeficiency virus type 1 integrase. *Antimicrob Agents Chemother* 42:140–146
- Menin B, Comino C, Moglia A et al (2010) Identification and mapping of genes related to caffeoylquinic acid synthesis in *Cynara cardunculus* L. *Plant Sc* 179:338–347
- Menin B, Comino C, Portis E et al (2012) Genetic mapping and characterization of the globe artichoke (+)-germacrene A synthase gene, encoding the first dedicated enzyme for biosynthesis of the bitter sesquiterpene lactone cynaropicrin. *Plant Sc* 190:1–8
- Miccadei S, Di Venere D, Cardinali A et al (2008) Antioxidative and apoptotic properties of polyphenolic extracts from edible part of artichoke (*Cynara scolymus* L.) on cultured rat hepatocytes and on human hepatoma cells. *Nutr Cancer* 60:276–283
- Miller M, Dunham J, Amores A et al (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Res* 17:240–248
- Moglia A, Comino C, Portis E et al (2009) Isolation and mapping of a C3'H gene (CYP98A49) from globe artichoke, and its expression upon UV-C stress. *Plant Cell Rep* 28:963–974
- Motzo R, Deidda M (1993) Anther and ovule culture in globe artichoke. *J Genet Breed* 47:263–266
- Niggeweg R, Michael A, Martin C (2004) Engineering plants with increased levels of the antioxidant chlorogenic acid. *Nat Biotech* 22:746–754
- Palmer LE, Rabinowicz PD, O'Shaughnessy AL et al (2003) Maize genome sequencing by methylation filtrations. *Science* 302:2115–2117
- Penalver R, Duranvila N, Lopez MM (1994) Characterization and pathogenicity of bacteria from shoot tips of the globe artichoke (*Cynara scolymus*). *Ann Appl Biol* 125:501–513
- Perez-Garcia F, Adzet T, Canigual S (2000) Activity of artichoke leaf extract on reactive oxygen species in human leukocytes. *Free Radic Res* 33:661–665
- Pittlern M, Ernst E (1998) Artichoke leaf extract for serum cholesterol reduction. *Perfusion* 11:338–340
- Portis E, Acquadro A, Comino C et al (2005a) Genetic structure of island populations of wild cardoon (*Cynara cardunculus* L. var. *sylvestris* (Lamk) Fiori) detected by AFLPs and SSRs. *Plant Sc* 169:199–210
- Portis E, Barchi L, Acquadro A et al (2005b) Genetic diversity assessment in cultivated cardoon by AFLP (amplified fragment length polymorphism) and microsatellite markers. *Plant Breed* 124:299–304
- Portis E, Mauromicale G, Barchi L et al (2005c) Population structure and genetic variation in autochthonous globe artichoke germplasm from Sicily Island. *Plant Sci* 168:1591–1598
- Portis E, Mauromicale G, Mauro R et al (2009a) Construction of a reference molecular linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus*). *Theor Appl Genet* 120:59–70
- Portis E, Acquadro A, Scaglione D et al (2009b). Construction of an SSR-based linkage map for *Cynara cardunculus*. *Proceeding of the 8th Plant Genomics European Meeting*, p 142
- Portis E, Acquadro A, Longo A, Mauro R, Mauromicale G, Lanteri S (2010) Potentiality of *Cynara cardunculus* L. as energy crop. *J Biotech* 150:S165–166

- Portis E, Scaglione D, Acquadro A et al (2012) Genetic mapping and identification of QTL for earliness in the globe artichoke/cultivated cardoon complex. *BMC Res Notes* 5:252
- Robba L, Carine M, Russell S, Raimondo F (2005) The monophyly and evolution of *Cynara* L. (Asteraceae) *sensu lato*: evidence from the internal transcribed spacer region of nrDNA. *Plant Syst Evol* 253:53–64
- Rondanelli M, Giacosa A, Orsini F et al (2011) Appetite control and glycaemia reduction in overweight subjects treated with a combination of two highly standardized extracts from *Phaseolus vulgaris* and *Cynara scolymus*. *Phytother Res* 25:1275–1282
- Rottenberg A, Zohary D, Nevo E (1996) Isozyme relationships between cultivated artichoke and the wild relatives. *Gen Res Crop Evol* 43:59–62
- Rottenberg A, Zohary D (2005) Wild genetic resources of cultivated artichoke. *Acta Horti* 681:307–311
- Scaglione D, Acquadro A, Portis E et al (2009) Ontology and diversity of transcript-associated microsatellites mined from a globe artichoke EST database. *BMC Genomics* 10:454
- Scaglione D, Acquadro A, Portis E et al (2012a) RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics* 13:3
- Scaglione D, Lanteri S, Acquadro A et al (2012b) Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. *Plant Biotech J* 10:956–969
- Schinor E, Salvador M, Ito I et al (2004) Trypanocidal and antimicrobial activities of *Moquinia kingii*. *Phytomedicine* 11:224–229
- Schneider G, Thiele K (1974) Die Verteilung des Bitter-stoffes Cynaropicrin in der Artischocke. *Planta Med* 26:174–183
- Sevcikova P, Glatz Z, Slanina J (2002) Analysis of artichoke (*Cynara cardunculus* L.) extract by means of micellar electrokinetic capillary chromatography. *Electrophoresis* 23:249–252
- Shao F, Merritt P, Bao Z et al (2002) A *Yersinia* effector and a *Pseudomonas* avirulence protein define a family of cysteine proteases functioning in bacterial pathogenesis. *Cell* 109:575–588
- Shimoda H, Ninomiya K, Nishida N et al (2003) Anti-hyperlipidemic Sesquiterpenes and new sesquiterpene glycosides from the leaves of artichoke (*Cynara scolymus* L.): structure requirement and mode of action. *Bioorg Med Chem Lett* 3:223–228
- Slanina J, Taborska E, Musil P (1993) Determination of cynarine in the decoctions of the artichoke (*Cynara cardunculus* L.) by the HPLC method. *Cesko-SloV Farm* 42:265–268
- Slanina J, Taborska E, Bochorakova H et al (2001) New and facile method of preparation of the anti-HIV-1 agent, 1,3-dicaffeoylquinic acid. *Tetrahedron Lett* 42:3383–3385
- Sousa MJ, Malcata FX (1996) Influence of pasteurization of milk and addition of starter cultures on protein breakdown in ovine cheeses manufactured with extracts from flowers of *Cynara cardunculus*. *Food Chem* 57:549–556
- Stamigna C, Saccardo F, Pandozy G et al (2005) In vitro mutagenesis of globe artichoke (cv. Romanesco). *Acta Hort* 681:403–410
- Timme R, Kuehl J, Boore J, Jansen R (2007) A comparative analysis of the *Lactuca* and *Helianthus* (Asteraceae) plastid genomes: Identification of divergent regions and categorization of shared repeats. *Am J Bot* 94:302–312
- Velasco R, Zharkikh A, Troggio M et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS one* 2:e1326
- Villegas R, Kojima M (1986) Purification and characterization of Hydroxycinnamoyl D-Glucose: quinate hydroxycinnamoyl transferase in the root of sweet potato, *Ipomoea batatas* Lam. *J Biol Chem* 261:8729–8733
- Wagenbreth D (1996) Evaluation of artichoke cultivars for growing and pharmaceutical use. *Beitr Zuchtungsforsch* 2:400–403
- Wang M, Simon J, Aviles I et al (2003) Analysis of antioxidative phenolic compounds in artichoke (*Cynara scolymus* L.). *J Agr Food Chem* 51:601–608
- Wiklund A (1992) The genus *Cynara* L. (Asteraceae, Cardueae). *Bot J Linn Soc* 109:75–123

- Yasukawa K, Matsubara H, Sano Y (2010) Inhibitory effect of the flowers of artichoke (*Cynara cardunculus*) on TPA-induced inflammation and tumor promotion in two-stage carcinogenesis in mouse skin. *J Nat Med* 64:388–391
- Ye S, Dhillon S, Ke X et al (2001) An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Res* 29:e88–8

Chapter 23

Genetic Diversity Assessment in European Cynara Collections

Mario Augusto Pagnotta and Arshiya Noorani

Contents

23.1	Introduction	560
23.1.1	Production and Importance	563
23.1.2	CYNARES: An EU Project	565
23.2	Diversity Assessment	566
23.2.1	Morphological Characterization	567
23.2.2	Biochemical Evaluation	567
23.2.3	Molecular Assessment	569
23.3	Diversity in European Collections	570
23.3.1	Variation Within and Among Cultivated Collections	570
23.3.2	CYNARES Collections	572
23.3.3	Variation Within and Among Cardoon Collections	576
23.3.4	DNA Conservation	577
23.4	Conclusions	578
	References	579

Abstract *Cynara cardunculus*, particularly artichoke, is an important horticultural crop in Italy as well as in France and Spain. Agronomic management of this crop is challenging as traditional varieties typically consist of heterogeneous populations. In this context, the EU-funded CYNARES was developed and implemented, with the aim of conserving, characterizing, collecting and utilizing the potential *C. cardunculus* diversity. The project promoted, among other activities, the development of *in vitro* culture techniques to obtain healthy, genetically uniform and rapidly propagated. Overall *C. cardunculus* genetic variability was evaluated using morphological, molecular and biochemical analyses. Agromorphological descriptors, based on those developed by IPGRI and UPOV, were used to characterize accessions. Molecular

M. A. Pagnotta (✉)

Department of Science and Technologies for Agriculture,
Forestry, Nature and Energy (DAFNE), University of Tuscia,
Via San Camillo de Lellis, 01100 Viterbo, Italy
e-mail: pagnotta@unitus.it

A. Noorani

Plant Production and Protection Division, FAO,
Viale delle Terme di Caracalla, Rome 00153, Italy

markers were used mainly for varietal identification, assessment of population diversity and relatedness investigations. Biochemical evaluation has been undertaken evaluating the nutraceutical of artichoke utilization. The EU project also established a *C. cardunculus* DNA bank focusing on advances in genomics and gene discovery.

Keywords Artichoke · Cardoon · Descriptors · Nutraceuticals · Molecular markers

23.1 Introduction

Cynara spp. belongs to the *Asteraceae* family which is composed of ten perennial and herbaceous species. *Cynara* spp. height ranges from approximately 0.50 m to 2 m with leaf segments ending with a spine in all species. *Asteraceae* flowers possess a specialized *capitulum*, technically called a *calathid* or *calathidium*, but generally referred to as a *flower head* and is formed by an involucre of many bracts. The bracts are free, and arranged in several rows, overlapping like the tiles of a roof. The flower head is constituted by numerous, generally violet, florets. The florets have five petals fused at the base to form a corolla tube. The calyx of the florets is always modified into a pappus of two or more teeth, scales or bristles and this is often involved in seed dispersion. There are usually five stamens, with the filaments fused to the corolla, while the anthers are connate. Pollen is released inside the tube and is collected around the growing style, expelled with a sort of pump mechanism or a brush. The mature seeds usually have little endosperm. Although there are two fused carpels, there is only one locule, and only one seed per achene-like fruit is formed.

Cynara species are diploid with $2n = 2 \times = 34$, clustered into two groups: one comprising seven species: *Cynara algarbiensis*, *Cynara baetica*, *Cynara cornigera*, *Cynara cyrenaica*, *Cynara syriaca*, *Cynara auranitica* and *Cynara humilis*, the other the complex *Cynara cardunculus* which includes the cultivated forms (Rottenberg and Zohary 1996; Fig. 23.1).

The wild cardoon (*Cynara cardunculus* var. *sylvestris* (Lamk) Fiori) (Fig. 23.2) is the progenitor of the two cultivated subspecies: the artichoke (*Cynara cardunculus* var. *scolymus* (L.) Fiori) and the cultivated cardoon (*Cynara cardunculus* var. *altilis* DC) (Sonnante et al. 2007; Table 23.1). All three subspecies are inter-fertile forming fully fertile hybrids, although genetic studies of wild cardoon have shown them to be more closely related to the cultivated cardoon than to the artichoke (Sonnante et al. 2002, 2004).

The species distribution ranges from Cyprus in the east to Portugal and the Canary Islands in the west. Its flowers have been used for centuries in the Iberian Peninsula for the manufacture of ovine and/or caprine milk cheeses (Sousa and Malcata 1997; 1998). The small and thorny capitula are gathered and sometimes sold in local markets in Sicily. The plant has been utilized as: (i) food (typical of the Mediterranean diet),

Fig. 23.1 Relationships among *Cynara* species and within *C. cardunculus*. Adapted by Sonnante et al. (2008) and Rottenberg and Zohary (1996) data

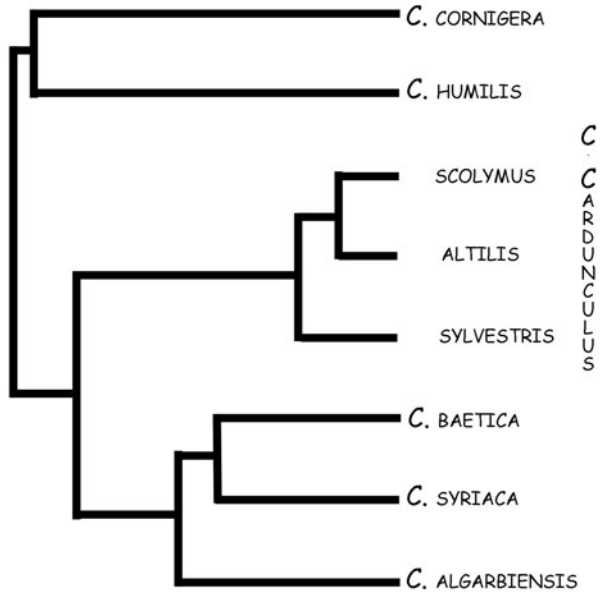


Fig. 23.2 Wild cardoon. (Photo A. Noorani)



Table 23.1 *Cynara* botanical classification

Kingdom	<i>Plantae</i> —Plants
Subkingdom	<i>Tracheobionta</i> —Vascular plants
Superdivision	<i>Spermatophyta</i> —Seed plants
Division	<i>Magnoliophyta</i> —Flowering plants
Class	<i>Magnoliopsida</i> —Dicotyledons
Subclass	<i>Asteridae</i>
Order	<i>Asterales</i>
Family	<i>Asteraceae</i> —Aster family
Genus	<i>Cynara</i> L.—cynara
Species	<i>Cynara cardunculus</i> L.—the wild cardoon, the cultivated cardoon and the artichoke

Fig. 23.3 The typical growth forms of artichoke **a** shorter with large capitula, and the cultivated cardoon **b** larger stalks and smaller capitula



(ii) lignocellulosic biomass for energy and paper pulp, (iii) seed oil for biodiesel fuel production (iv) beverage, (v) nutraceutical purposes (Raccuia and Melilli 2004).

Current research has shown several health benefits, including promotion of blood circulation, induction of choleresis, inhibition of cholesterol biosynthesis and antioxidant effects (Fратиanni et al. 2007; Lattanzio et al. 2009; Lombardo et al. 2010; Pandino et al. 2010, 2011a; Bonasia et al. 2010).

Globe artichoke is usually vegetatively propagated by means of *carducci* (basal shoots) or *ovoli* (semi-dormant shoots with a limited root system) (Snyder 1979), while cultivated cardoon is raised from seed and cropped as an annual plant. The two crops are thought to be the result of human selection pressure for either large, non-spiny heads or non-spiny, large stalked tender leaves (Basnizki and Zohary 1994; see Fig. 23.3). Sonnante et al. (2004) considered the two cultivated forms to be the result of concurrent directional selection for distinct traits, and not disruptive selection.

The species is characterized by a high degree of heterozygosity (Mauromicale and Ierna 2000). Studies based on variation of isozymes and molecular markers such as RAPDs and AFLPs (Rottenberg et al. 1996; Sonnante et al. 2002, 2004; Lanteri et al. 2004a; Raccuia et al. 2004) have confirmed that both crops evolved from the wild cardoon gene pool, thereby confirming that it is the progenitor of both cultivated types. AFLP marker studies have shown that all *C. cardunculus* samples share a high genetic similarity compared with the other *Cynara* wild species. Despite this, artichoke germplasm is well separated from both wild and cultivated cardoon samples (Sonnante et al. 2004).

Further studies based on rDNA spacer sequences, (Robba et al. 2005; Sonnante et al. 2007), show close agreement with the phylogeny proposed by Wiklund (1992). The analyses showed that the leafy cardoon is genetically closer to wild germplasm of Spain rather than wild germplasm of Italy or Greece (Sonnante et al. 2007). Pignone and Sonnante (2004) have hypothesized that the artichoke was possibly domesticated in Sicily, while cardoon originated in the western range of the Mediterranean, probably within Spain or France (Sonnante et al. 2007). Evidence indicates that the

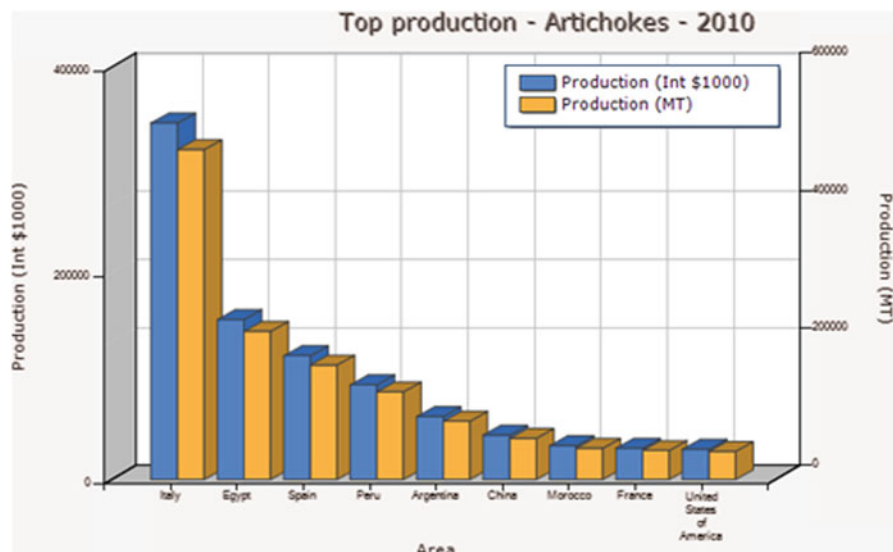


Fig. 23.4 Global production data for artichoke (FAOSTAT 2009)

domestication of artichoke took place around the beginning of the first millennium (Foury 1989; Pignone and Sonnante 2004) while domestication of cardoon took place in the first half of the second millennium. These studies revealed that the evolution of the whole genus took place as recently as 20,000 years ago (Fig. 23.1). The domestication of artichoke and cardoon are now believed to have been two distinctive events, separated in time and in space, which led to the two crops diverging in propagation and end use.

23.1.1 Production and Importance

Cynara cardunculus in general and artichoke in particular, is the third most important horticultural crop in Italy after potato and tomato (see Fig. 23.4). Native to the Mediterranean area, this is where about 90 % of global production is found (FAOSTAT 2009, <http://faostat.fao.org>).

Globe artichoke (*Cynara cardunculus* var. *scolymus* L.) is native to Southern Europe around the Mediterranean basin and North-Western Africa (Bianco 2005). Southern Italy and Sicily have been recently considered as the origin of its domestication (Pignone and Sonnante 2004). The origin of the artichoke is often associated with the Arabs, who were prevalent in the southern Mediterranean during Medieval times (Idrisi 2005). Due to its long-term cultivation in Italy, it is here that the diversity of globe artichoke autochthonous germplasm is greatest, i.e. the “primary

cultivated gene pool” (Pignone and Sonnante 2004). The artichoke is therefore well adapted to the different pedoclimatic conditions present in Italy and other parts of the Mediterranean (Cravero et al. 2010).

The globe artichoke sector is currently undergoing an economic crisis due to various contributing factors, notably the arrival of products from abroad that ensure availability throughout the summer. In addition, the European Union has not yet adopted effective measures to regulate the entry in the area of non-European products and their negative influence on the EU market prices (Agostinucci and Loseby 2007). Farmers’ profits are also affected by the high cost of labour required for crop cultivation and harvesting.

The agronomic management of the crop is challenging as traditional varieties typically consist of heterogeneous populations. Artichoke varieties are classified based on characteristics of the capitula, harvest times (early or late) and distinct clonal varietal groups. Early varieties mature typically around autumn-winter of the southern Mediterranean, and continue to produce until spring, while the later maturing varieties, represented by spring cultivars (e.g. Romanesco), provide products from February to April (Crinò et al. 2008). The nomenclature of both early and spring varieties is not always very clear and there are cases of homonyms. Additionally, cultivars are often labelled based on the name of their area of cultivation, leaving much room for overlap and error.

The effectiveness of vegetative reproduction is limited due to the low rate of multiplication and high transmission rate of pathogens. The development of *in vitro* culture techniques to obtain healthy, genetically uniform and rapidly propagated material is a viable alternative to the standard system of propagation. If conventional techniques are not available, e.g. when seed conservation is not be feasible and field genebanks may be considered costly or risky, *in vitro* techniques are used to complement conservation and utilization of genetic diversity.

Over the last ten years, *in vitro* storage of cultures has been applied to a wide range of species and culture systems with varying degrees of success. Successful slow growth methodologies have been developed for different species (Withers et al. 1990; Engelmann 1991; Janeiro et al. 1995). Growth reduction is obtained by (i) growth limitations of differentiated *in vitro* microplantlets and developing meristem cultures, or (ii) suspension of growth and metabolism by employing ultra-low temperatures (Sarkar and Naik 1998). Thus, for example, the “Romanesco” globe artichoke C3, resulting from a clonal population of a traditional population ‘Castellammare’, is now propagated by *in vitro* culture. The variety ‘Terom’ has been derived from Tuscany Violet by the same procedure (Soressi 2003). Tissue culture is regularly used for micropropagation and production of disease-free plantlets of spring genotypes (Castiglione et al. 2007). Moreover, new seed propagated F₁ hybrids (e.g. Madrigal, Concerto, Opal, and Tema) are now increasingly cultivated (Zaniboni 2009; Bonasia et al. 2010). Conversely, the introduction of foreign F₁ hybrids, as well as the use of Italian spring micropropagated clones as a response to market demand, represents a risk of genetic erosion of autochthonous germplasm.

23.1.1.1 Artichoke Hybrids

To overcome the problem and the costs of *in vitro* micropropagation, F₁ globe artichoke hybrids are now being used widely. Further interest and development of the F₁ hybrid program was taken by the private sector (Foury et al. 2005). In fact, hybrids have several advantages over traditional cultivars such as: (i) labour saving and cheap operation due to mechanical seed sowing; (ii) conversion of globe artichoke into an annually grown crop and introduction into crop rotations; (iii) efficient use of both water and fertilizers; (iv) increased resistance to pathogens (mainly viruses) and pests; (v) greater potential for organic cultivation due to vigorous and healthy growth despite low chemical inputs; (vi) facilities for plant production nurseries. However, there are issues regarding hybrid cultivation, including: (i) correct management of the production cycle for seed propagated plants, (ii) the adaptability of the F₁ hybrids to diverse environments as well as to production and commercialization calendars, (iii) the difficulty of producing high seed yields and hence economical production of F₁ hybrid seed.

During the last 15 years, a program of genetic improvement was carried out at Tuscia University (Italy) in order to obtain new F₁ hybrids stable and with characteristics of interest for Italian market (Saccardo 2009; Lo Bianco et al. 2012). The new hybrids were developed using sterile male genotypes since stable F₁ hybrids need to be obtained from stable parent lines. A first selection was run on male sterile clones; some stable male fertile genotypes were also selected. Different cross-combinations were developed both in Italy and in USA in order to test the different combining abilities and to evaluate the different hybrids in different environmental areas. Due to the environmental conditions present in California, seed production was higher and more homogeneous. A major focus of the program consisted in finding an evaluation system capable of distinguishing which hybrids were more homogeneous than others. Some F₁ hybrids have now been registered and will be soon ready for commercial production.

23.1.2 CYNARES: An EU Project

The characterization and conservation of *Cynara cardunculus* germplasm is now an international concern. Addressing this, the European Commission (Directorate-General for Agriculture and Rural Development under Council Regulation (EC) No 870/2004) sponsored the European Project 'CYNARES'. The project (2008–2012) has seven partners from France, Spain and Italy who share European germplasm collections which have been assessed at the morphological, biochemical and molecular levels as well as for disease resistance. The conservation of *C. cardunculus* germplasm has been undertaken based on the Convention on Biological Diversity, the FAO's Global Plan of Action for Plant Genetic Resources for Food and Agriculture, and the International Treaty on Plant Genetic Resources for Food and Agriculture.

France, Italy and Spain were key project partners as traditional artichoke cultivars are grown predominantly in these Mediterranean countries. These include 'Petit Violet de Provence', 'Catanese', 'Spinoso Sardo', 'Romanesco' and 'Blanca de Tudela'. In Spain a single ecotype, 'Blanca de Tudela', represents 90 % of the Spanish production. In France, no more than five or six clones are commonly cultivated; almost all current French production is based on the globe artichokes 'Camus de Bretagne', 'Castel' and the small cylinder artichoke 'Petit Violet de Provence'. In Italy only 'Violetto di Sicilia', 'Romanesco' (mainly the clone C3), and 'Spinoso Sardo' are cultivated over large areas. The low diversity of cultivated globe artichoke germplasm, as compared to the large diversity originally present, is an indicator of the fast pace of genetic erosion.

Compounding the issue of low levels of diversity is the difficulty in defining *C. cardunculus* varieties. This is due to years of cultivation in various geographical sites where each cultivar has been locally named according to the place where it was cultivated (Bianco 1990). Thus, the accessions need to be rationalized by improving core collections and avoiding duplication (Frankel 1984; Brown 1989a, b).

CYNARES is a cooperative project which focuses on artichoke as it is a commercially more important crop than is cardoon. In order to protect artichoke biodiversity present in Europe and to encourage traditional foods, a few lines now have been registered as Protected Geographical Indication (PGI) products. These are the Italian 'Romanesco' globe artichoke (CE Reg. no. 2066/2002 of the Commission, published by GUCE no. L 318 of 22/11/2002) and 'Tondo di Paestum' (CE Reg no 2081/92, published Official Journal C 153, 01/07/2003) and Spanish 'Blanca de Tudela' (CE Reg no 1971/2001 of the Commission, 09/10/2001).

Generations of breeding have given rise to artichoke varieties that are optimally adapted to their specific regional environmental conditions, but intensive agriculture and material dispersal increasingly threaten this diversity. Another element is the loss of traditional knowledge since artichoke production requires skilled labour. Further, the project promotes efficient artichoke diversity management, while creating new varieties tailored to consumer requirements. To do this, assessment of nutritional and pharmaceutical value of artichoke through biochemical characterization has been undertaken. Consequently, the project promotes the sustainable use of *Cynara cardunculus* germplasm while providing a better cultivar choice for the consumer. Screening disease-resistant genotypes is also a key activity of the CYNARES project with a view to ensuring environmentally friendly cultivation practices. The knowledge gained is disseminated to all partners (including deliverables), through the development of websites, booklets, symposia, and scientific/technical publications.

23.2 Diversity Assessment

Genetic variability is evaluated using morphological, molecular and biochemical analyses (Lahoz et al. 2011). Conservation strategies include protection of wild species, cultivated populations and traditional crop varieties on-farm, together with

their preservation in *ex situ* genebanks. To reorganize and utilize germplasm collections, it is essential to properly characterize the material for morphological, biochemical, and molecular traits as well as the extent and distribution of its genetic variation present in *in situ* and *ex situ* collections.

23.2.1 Morphological Characterization

Morphological characterization is now required as the first parameter of germplasm assessment, allowing selection based on traits of commercial value and environmental requirements. In order to assess diversity of the germplasm conserved in fields and in genebanks, evaluation in the context of the CYNARES project, taking into account the specific environmental requirements of *Cynara cardunculus*, was carried out. Descriptors, based on those developed by IPGRI and UPOV, were used for identifying accessions (see Table 23.2 for a list of the descriptors used).

Most of the *Cynara cardunculus* characterization studies utilize the agromorphological traits indicated in the UPOV descriptor list, including: plant height, weight, date of flowering of the central flower head, leaf length, and head shape (Noorani et al. 2012a). Montemurro et al. (2012) found that bracts colour and head shape are useful traits in classification of their collections. Cadinu et al. (2012) subdivided the collection looking at genotype precocity (late versus early) within each of the four main varietal types (Catanese, Spinoso, Romanesco and Violetto) (Porceddu et al. 1976), taking into consideration head shape, colour and presence/absence of spines.

23.2.2 Biochemical Evaluation

Biochemical use of artichoke is not restricted to its edible parts, i.e. to immature flower heads commonly used as fresh, frozen or canned delicacies (Bianco 1990), but also to the leaves, stem and roots that are utilized as sources of (i) forage for livestock, (ii) inulin, (iii) feedstock for the extraction of secondary metabolites (cynarine, luteolin, chlorogenic acid, cynaroside), and (iv) alcoholic beverages. Since ancient times, the leaves have also been widely used as hepatoprotectors and choleric agents in herbal medicine (Bruneton 1995; Gebhardt 2002). Artichoke contains polyphenolic derivatives (mono- and di-caffeoylquinic acids and flavonoids) showing strong antioxidant, choleric, hypocholesterolemic and hepatoprotective effects. It also contains sugars and the fructan inulin, which has an important role in the human diet as prebiotic soluble fibre and contribute to make the artichoke a “functional food”.

In Germany and China, artichoke is used in pharmaceutical products for its high content of polyphenols such as caffeoylquinic acids (total range from 2–8 %), which are considered to be part of the active principle involved in antioxidant and choleresis (stimulation of biliary secretions) activities. Flavonoids (total range from

Table 23.2 Descriptors used in assessing morphological diversity

<i>Plant</i>	
Height (including central flower head)	Number of lateral shoots on main stem.
<i>Main Stem</i>	
Height (excluding central flower head)	Diameter (at about 10 cm below central flower head)
Distance between the central flower head and the youngest well developed leaf	
<i>Leaf</i>	
Attitude	Long spines
Lobbing/incisions	Length
Lobe: shape of tip(excluding terminal lobe)	Number of lobes
Lobe: number of secondary lobes	Lobe: shape of tip in secondary lobes.
<i>Leaf Blade</i>	
Intensity of green colour (upper side)	Intensity of hairiness on upper side
Hue of green colour Intensity of grey hue	Leaf blade: blistering
<i>Petiole</i>	
Anthocyanin coloration at base	
<i>Central Flower Head</i>	
Length	Time of appearance
Diameter	Anthocyanin coloration of inner bracts
Size	Density of inner bracts
Shape in longitudinal section	
Shape of tip	
<i>First Flower Head on Lateral Shoot</i>	
Length	Shape in longitudinal section
Diameter	
<i>Outer Bract</i>	
1st Head	Colour (external side)
Length of base	Hue of secondary colour
Width of base	Reflexing tip
Thickness at base	Size of spine
Main shape	Depth of emargination
Shape of apex	Mucron.
<i>Receptacle</i>	
Diameter	Shape in longitudinal section
Thickness	
<i>Productivity</i>	
Weight of the central flower head	Total number of flower heads
Weight of the secondary flower heads	Total weight of flower heads excluding central and secondary heads
Harvesting date of the secondary flower heads	

0.35–1.35 %) have anti-inflammatory and antioxidant activities. It might act anti-cholestatically at least in some types of intra-hepatic cholestasis. Inulin, a fructan complex, also plays an important nutritional role.

Several studies have conducted to characterize biochemically different globe artichoke genotypes, to evaluate genetic variability existing in *Cynara* spp. germplasm

and to select the most suitable globe artichoke genotypes for fresh consumption or/and industrial processing (Fратиани et al. 2007; Pandino et al. 2010, 2011a, b; Lombardo et al. 2010; Bonasia et al. 2010). The biochemical compounds in the germplasm have shown the existence of great variability for most compounds. This variability is not randomly distributed, but follows the geographic origin pathway; for example, chlorogenic acid is higher in the Violetto di Sicilia clones, followed by spring types of Violetto di Provenza (Melilli et al. 2007). Other possible uses of *Cynara* spp. have also been considered over the last few years such as (i) seeds for oil (Foti et al. 1999; Curt et al. 2002; Raccuia and Melilli 2007; Raccuia et al. 2011), (ii) roots for inulin (Raccuia and Melilli 2004, 2010), (iii) biomass for energy (Raccuia and Melilli 2007; Angelini et al. 2009; Ierna and Mauromicale 2010; Gominho et al. 2011), (iv) fibre for pulp and paper industry (Antunes et al. 2000; Gominho et al. 2001, 2009), (v) green forage for ruminant feeding (Fernández et al. 2006), (vi) natural rennet for traditional cheese making (Fernández et al. 2006; Galán et al. 2008), and (vii) plant for metal-accumulation (Hernández-Allica et al. 2008). These applications of the crop are linked principally to European Union research support on new agricultural by-products (industrial raw materials) and has led to increasing interest in aboveground globe artichoke biomass.

This interest is principally due to the easy adaption of the crop to various Mediterranean climates, characterized by low annual rainfalls and hot dry summers (Fernández et al. 2006), and to the relatively low crop energy input and the large biomass productivity (Angelini et al. 2009). Until now, several studies on globe artichoke as energy crop existed but there was little work on the use of its biomass as a raw industrial material for the recovery of phenolic active compounds. Specially, there was a lack of useful data on active biocompound extraction from globe artichoke biomass for the pharmaceutical industry to meet the increasing demand for natural antioxidants due both to health concerns linked to the use of synthetic antioxidants such as BHT and BHA, and to consumers' preferences (Llorach et al. 2002).

In this framework, a sustainable large-scale production of biomass for biocompound extraction was evaluated and the possibility of using crop biomass without upsetting traditional agricultural practices while increasing farmers' income was considered.

23.2.3 Molecular Assessment

Advances in genomics have provided technologies for high throughput analyses of plant genomes with potential use in gene discovery and germplasm collections. The establishment of DNA banks facilitates this screening of DNA from large numbers of plant accessions (Rice et al. 2006). Markers at the DNA level represent the ideal tool since they characterize individuals directly at the level of genotype, without any concern on the effect of environment, plant tissue and developmental stage and, furthermore, their number is virtually infinite (Mondini et al. 2009). Molecular markers

can be used to identify populations and landraces, clustering populations and genotypes, as well as genotype fingerprinting. In the last decade several molecular markers have been utilized in *Cynara cardunculus* populations, such as AFLP (Lanteri et al. 2004a; Pagnotta et al. 2004; Raccuia et al. 2004; Portis et al. 2005; Mauro et al. 2009; Acquadro et al. 2010), RAPD (Lanteri et al. 2001; Sonnante et al. 2002), ISSR (Crinò et al. 2008; Lo Bianco et al. 2011) and SSR (Acquadro et al. 2005; Sonnante et al. 2008; Mauro et al. 2009). The molecular markers were used mainly for variety identification and genetic diversity and relatedness investigations (Tivang et al. 1996; Sonnante et al. 2002, 2008; Crinò et al. 2008), plant breeding as selection tools and to characterize parental lines in F1 hybrid constitution (Lo Bianco et al. 2011), as well as to identify the most probable progenitors of a given hybrid (Messmer et al. 2002). Furthermore, genetic linkage maps have also been recently developed in artichoke (Lanteri et al. 2006; Lanteri and Portis 2008; Acquadro et al. 2009; Portis et al. 2009; Sonnante et al. 2011b). The first two were from the same mapping population obtained by crossing two artichoke types, the other (Portis et al. 2009) was obtained by using an F1 population derived from a cross between artichoke and cultivated cardoon, while the last (Sonnante et al. 2011b) utilized a cross between artichoke and wild cardoon. Also, some specific genes are reported in the linkage maps, such as genes involved in the synthesis of phenylpropanoid compounds (De Paolis et al. 2008; Portis et al. 2009; Menin et al. 2010; Sonnante et al. 2010).

The crosses between artichoke and both cultivated and wild cardoon resulted in an explosion of variability due to the heterozygous nature of the species, which results not only in a great diversity, useful to map the markers, but also in new morphological types which could be useful in selection activities and especially for ornamental purposes (Lanteri et al. 2012).

23.3 Diversity in European Collections

23.3.1 Variation Within and Among Cultivated Collections

23.3.1.1 The Romanesco Artichoke

Varieties of 'Romanesco' globe artichoke from different central Italian areas, cultivated for centuries by local farmers and thus a good example of *in situ*-selection (Portis et al. 2005), have been evaluated by molecular and morphological means. Their value lies in the preservation of traditional genetic resources and their provision of genotypes/genes useful for the production of new materials suited to market requirements. Some of the studies aim at (i) characterizing these landraces in terms of genetic and molecular profiles; (ii) investigating the genetic variability existing within and among landraces; (iii) identifying genetic resources for the development of future breeding programs. Landraces were clustered on the basis of morphological and molecular traits (Crinò et al. 2008; Ciancolini et al. 2012). These clusters

agreed with the landraces' original area even if no significant correlation between genetic similarity and morphological trait matrices was revealed by the Mantel test. The mismatches between genetic distance and geographic origin within a globe artichoke typology are common also in typologies other than Romanesco (Sonnante et al. 2002; Lanteri et al. 2004b; Portis et al. 2005). A probable cause of this discrepancy could be that the morphological traits are based on a small number of genes, whereas the molecular ones are based on several loci widely distributed across the genome. The degree of morphological plasticity expressed by globe artichoke increases the value of DNA markers since they are unaffected by variation in either the environment or the developmental stage. The complementarity of both morphological and molecular analyses can contribute to the accurate identification and classification of landraces in the management of plant genetic resources for *in situ* and *ex situ* conservation as well as for use in plant breeding programs (Crinò et al. 2008; Lo Bianco et al. 2012).

Information obtained by genetic characterization is important to classify and identify the correct germplasm to be conserved and used in breeding activities. The selection of clones within local landraces is a fundamental pre-requisite for the release of new varieties, or for the development of homogeneous lines useful as parental hybrid combinations in plant breeding programs. Preserving traditional globe artichoke germplasm is important for safeguarding against the risk of genetic erosion. This opens new perspectives for the genetic improvement of globe artichoke, which up to now has been neglected. Characterization activity carried out on Romanesco landraces collected in Latium region (Ciancolini et al. 2012), has resulted in the registration of three globe artichoke clones (Ancora et al. 2012).

The number of private bands (i.e. fragments present in only one accession) found by Crinò et al. (2008) in 'Romanesco' clones is comparable to that observed by Lanteri et al. (2004b) and Sonnante et al. (2004, 2007) in collections with all the global artichoke types. The predicted heterozygosity was different (0.37 versus 0.67) from that estimated in wild cardoon using SSR markers (Portis et al. 2005), probably because SSR loci are associated with a high number of detectable alleles, whereas AFLP and ISSR loci are dominant.

23.3.1.2 The Spinoso Artichoke

Landraces of Spinoso from Sardinia and Sicily have been evaluated by RAPD (Lanteri et al. 2001) and AFLP (Portis et al. 2005) respectively. The AMOVA (Analysis of Molecular VARIance) highlighted the existence of diversity between populations (28.1 % of the total genetic diversity), but the majority was due to differences within the population (71.8 %). This demonstrates the possibility of carrying out clonal selection within a population. The comparison between several clones (Muntoni and Poddie 2002) resulted in the identification of biotypes distinguished by their agronomic and commercial characteristics. *In vitro* culture of those biotypes showed that such positive characteristics were maintained. Out of 80 clones, 9 were selected and characterized for agronomic and biometric parameters as well as yield stability (Mallica et al. 2004). The clones 110/14 and 108/11 revealed a high yield potential;

while clone 39 resulted in the earliest yield. The selected clones of Spinoso Sardo could represent a starting point for recommended nursery activity in the future. The genetic distribution detected (Portis et al. 2005) supplies important information for the implementation of ‘on farm’ germplasm preservation strategies.

23.3.2 CYNARES Collections

In the framework of the CYNARES project, sponsored by the AGRI GEN RES Community Programme (European Commission, Directorate-General for Agriculture and Rural Development, under Council Regulation (EC) No 870/2004), Italian, French and Spanish partners collected and evaluation accessions of different *C. cardunculus* germplasm (Pagnotta 2011a).

23.3.2.1 Central Italian Collections at University of Tuscia and ENEA

ENEA and the University of Tuscia (Crinò et al. 2011) jointly hold 31 Romanesco accessions of globe artichoke clones, four spring landraces from Marche region (Montelupone A, Montelupone B, Jesino, Ascolano), one spring landrace from Tuscany region (Pisa), and two spring landraces from Campania region (Bianco di Pertosa, Tondo rosso di Paestum). Offshoots of the Romanesco clones were collected from morphologically distinct plants in farmer fields around Rome, where the Romanesco landraces ‘Campagnano’ and ‘Castellammare’ have been traditionally cultivated. *In vitro* multiplication of shoot apices was performed at ENEA; micro-propagated clones derived from selected mother plants were matured in a greenhouse, before being transferred to experimental fields in Cerveteri (Rome, Italy). Offshoots from the eight central and southern Italy landraces (Montelupone A, Montelupone B, Jesino, Ascolano, Pisa, Pertosa, Paestum, Tondo Rosso di Paestum) were used for growing plants in the catalogue field of Cerveteri.

Most clones present in the Cerveteri experimental field are also in *C. cardunculus* *in vitro* collections (10 copies/clone) held at ENEA. Morphological characterization of all materials in the collection was conducted at the Latium Regional Agency for the Development and the Innovation of Agriculture (ARSIAL) in Cerveteri. From each clone, morphological data were recorded twice weekly during the period of the most rapid growth, and weekly during the remainder of the period, on three representative plants/clone for each of the four repetitions and landraces during the project (2007–2011). Each individual plant was phenotyped using descriptors jointly defined by CYNARES partners (see sect. 2.1).

Seven accessions of cultivated and one of wild, cardoon are also included in the ENEA/University of Tuscia collections of *Cynara* spp. These accessions were collected from natural populations, open pollinated and evaluated in experimental fields in Tarquinia (Viterbo, Italy). In the framework of CYNARES project, selfings of three plants were selected from each accession. Other cardoon accessions originating from

international genebanks such as IPK Gatersleben, Germany, and VIR, Russia, were also characterized at the University of Tuscia's experimental fields (Viterbo).

Morphological characterization for vegetable and biomass production was made and accomplished on both globe artichoke and cardoon accessions. All data related to morphological characterization both for vegetable and biomass production were statistically analysed. Based on the morphological traits recorded for all genotypes, similarity dendrograms were constructed using the agglomerative hierarchical cluster analysis. Clones 22, 23, and Grato 1 along with Pisa, Tondo Rosso di Paestum, Bianco di Pertosa, and Ascolano showed the highest values of flower head production in terms both of weight and number. The clones 18, Campagnano, 2, 11, and 5 along with Ascolano, Bianco di Pertosa, and Pisa were selected for the highest biomass production.

The same morphological characterization carried out on cardoon accessions revealed that AFM, AFN, AFFGA, AFB, and AFS were the most productive genotypes for number of total flower heads. All the cardoon accessions analysed, other than except AFS showed levels of high biomass production.

Finally, the biomass of the most significant clones of globe artichoke and cardoon were characterized for antioxidant content (total polyphenols, cynarin, caffeoylquinic acids, chlorogenic acid, inulin, luteolin and apigenin), and the yield of useful bio-compound that could be extracted from cardoon. It is interesting to note that cardoon has been never selected for these traits, hence the possibility for breeding activities could be very promising (Ciancolini et al. 2013).

Cardoon accessions have been also tested for tolerance to *Verticillium dahliae*. Thirty plants per clone of Italian artichoke and cardoon were laid down in a randomised blocks design with three replications and inoculated with Italian collections of *V. dahliae* inocula. A control of non-inoculated plants was also included. The utilized inoculum was prepared by blending 15 day-old colonies grown on potato-dextrose agar (PDA) in tap water; a final concentration of 10^6 conidia/ml. Inoculations were made by dipping the basal part of the offshoots into the inoculum immediately before transplanting. The percentage of diseased plants and the disease severity index were visually estimated on the basis of the external symptoms (foliar yellowing, necrosis and plant stunting using a quantitative 0–6 scale), and the discoloration in the main stem of each plant by cutting it transversely at soil level (using a 0–5 scale). The results highlight the existence of clones with different degree of resistance to *Verticillium* (Pagnotta et al. 2012)

23.3.2.2 Spanish Collections at Cartagena University

The artichoke and cardoon accessions evaluated by Universidad Politécnica de Cartagena (UPCT) (Egea-Gilabert et al. 2011) were donated by Instituto Técnico y de Gestión Agrícola (ITGA) from Navarra. All the accessions were grown at the “Tomás Ferro” Experimental Agro-Food Station of UPCT (37° 41' N; 0° 57' W).

Ten artichoke stumps (five Spanish Blanca de Tudela-type: INIA-D, INIA-B, ITGA, Clon 303 and Cabeza de Gato; two Italian: Spinoso Sardo and Moretto; and

three French: Salambo, Violet de Provence and Macau) were grown and characterized according to UPOV descriptors over three seasons (2007–2008, 2008–2009 and 2009–2010). The results showed that the number of flower heads per plant was variable among cultivars, with an average of all cultivars of 13.2, 13.8 and 9.9 flower heads for the first, second and third season, respectively. Overall, Clon 303 produced the largest number of flower heads in the trial, while Macau produced the lowest. The heaviest flower heads were produced by Salambo and Cabeza de Gato in the first and second seasons and by Salambo in the third one. In general, the average weight of all cultivars was increased in successive seasons. Regarding the diameter of the flower heads, the behaviour of the cultivars differed over the seasons. On average, the second season was the most productive, followed by the third. The highest yields were obtained with clones of Blanca de Tudela, INIA-B and Clon-303 in the first season, Salambo in the second and Spinoso Sardo in the third. The cultivar with the lowest productivity was Macau in the three seasons. Cabeza de gato also had a low yield. When analysing the cumulative monthly production, the results showed that the earliest cultivars in all three seasons were the clones of Blanca de Tudela, which began production in November in the first and second season and in October in the third one. By contrast, the latest cultivars were Macau, Moretto and Salambo, which began their production in March in the three seasons. Additionally, a postharvest study was done focussed in Salambo, Cabeza de Gato, Macau and Moretto. Artichokes were harvested at commercial maturity and were prepared both as a minimally processed product and a microwave-processed, with two kinds of presentation: hearts and halves hearts. The results showed that Macao had the highest respiration rate and Salambo was the least susceptible to browning. The microwave treatment was effective in maintaining colour and sensory quality while it progressively increased the total phenolic content and the total antioxidant capacity for Macau and Moretto, without significant changes for Salambo.

Ten cultivated cardoons (Blanco de Valencia, Blanco de Huerva, Sarramián, Verde Calahorra, Blanco Peralta, Del Cortijo, Verde de Huerva, Llano de Espaa, Rojo de Agreda and Verde Peralta) were transplanted at the beginning of August every year since 2007. All cardoon cultivars were harvested and characterized at the end of December. Four quantitative characteristics from the plant and six quantitative and three qualitative characteristics from the leaf stalk were measured. The most productive variety was Llano de Espaa, with a yield of 95.5 t ha⁻¹ and the least productive was Blanco de Huerva with 65.1 t ha⁻¹.

23.3.2.3 Spanish Collections at ITGA

ITGA, the Technical and Agricultural Management Institute of Navarra, performs tasks that fit in three different but perfectly coordinated blocks: research, experimentation on the field and diffusion of results through a network of agricultural advisors (Macua et al. 2011). ITGA evaluates and conserves 63 accessions of cardoon, 48 of which are from growers in different areas of Spain and from the Zaragoza Geoplasm Bank and the other are 11 from France and 4 from Italy.

In Spain, production of artichoke is based almost exclusively on the clonal propagated variety Blanca de Tudela. At present, within the most important artichoke production areas of Europe, the number of varieties has decreased to four: two of them for long cycle or early-maturing; Blanca de Tudela and Violet de Provence, and the other two of short cycle; Romanesco and Camus de Bretagne. Different clones of Blanca de Tudela, 14 French varieties and 5 from Italy, constituted the beginning of the ITGA collection and increased with mutations of Blanca de Tudela, Cabeza de Gato and Carderas. Subsequently, the collection has grown, with 59 varieties of vegetative propagation from Spain, Italy and France as of 2011.

Morphological characterization has been carried out on: ITGA selection, Cabeza Gato, INIA D, PAT-89 and Carderas, from Blanca de Tudela; Camus de Bretagne, Salanquet, Crysantheme, Camerys, Macau, France, Calico Rojo, Calico Verde, C-3, VP-41 and Hydes, from France; Apolo, Brindisi, Campagnano, Hysponos, Italiana, Masedu, Moretto and Mutación Romanesco, from Italy; and Criolla, from Argentina.

The morphological characterization of cardoon, based on descriptors developed by the project's partner teams, has been made for: 29 accessions from Spain (Acequi, Blanco de la Huerva, Blanco de la tierra, Blanco de Peralta, Blanco de Valencia, Blanco Francés de Valencia, C-001-Rosa de Agreda, C-002-Blanco, C-003-Cadrete, Cimbri, Del Cortijo-Arsenio, L'Horta, Lumbier, Llenu de Espaa, NC-072586-Segovia, NC-072684-Avila, NC-072783-Avila, NC-0722811-Avila, NC-076391-Valladolid, Penca Blanca Ancha, Rojo de Agreda, Rojo de Corella, Rojo Poeta, Sarramian, Verde de Calahorra, Verde de la Huerva, Verde de Tafalla, Verde Vicsar and Verde de Peralta), 3 from Italy (Bianco Avorio, Gigante de Romagna and Gobbo de Niza) and 5 from France (Blanc Ameliore, Plein Blanc Inerme, Rouge Dâlger, Vert de Vaulx Envelin and Epineux de Plainpalais).

23.3.2.4 Italian Collection at CNR Bari

The Institute of Plant Genetics of CNR, Bari, Italy holds a collection of *Cynara* samples (Sonnante et al. 2011a), mostly of *Cynara cardunculus*, including globe artichoke, a few accessions of cultivated cardoon and a number of wild cardoon populations from the Mediterranean basin. Within the framework of the EU CYNARES project, a subset of the globe artichoke collection was evaluated. Most accessions were directly collected from farmers' fields in southern Italy, especially in Apulia, where artichoke is widely cultivated. Some samples were exchanged between project partners.

Collected data were analysed statistically to assess qualitative and quantitative characteristics of the accessions. Frequencies were calculated for qualitative data, while quantitative data were analysed by descriptive statistics. A subset of the samples used for morpho-agronomic evaluation was employed for biochemical investigation that focused on the analysis of polyphenols and inulin in secondary flower heads. Polyphenol fractions evaluated were mainly caffeic acids, its derivatives (caffeoylquinic acids) and flavonoids. Different genotypes displayed different contents

of polyphenol fractions and results were reproducible. Another study was dedicated to the content of polyphenol fractions in flower head bracts and receptacles, and in leaves in different developmental stages of the globe artichoke plant.

23.3.2.5 French Collection at GEVES

All these accessions are listed in the National Collection of the CYNARA network, the French network dedicated to maintenance of 28 artichoke vegetatively propagated varieties (clones) and 5 open pollinated cardoon varieties (populations) (Jouy et al. 2008). Morphological evaluation, during the first and second years, was carried out in different sites in France. The data was validated by *a priori* pictures (central flower head) taken at the INRA Plougoulm station, in open fields and glasshouse trials, as well as at the INRA Ploudaniel station (greenhouse and open field). The number of varieties introduced in the French National Collection was optimized in order to limit the charge of maintenance over the long term. The preservation-regeneration tasks were performed thanks to the *in vivo* conservatories.

23.3.3 Variation Within and Among Cardoon Collections

Analyses in cardoon are limited in number compared with the ones utilized in artichoke, with morphological traits in cardoon taken from the UPOV list. However, principal component analysis showed that a reduced number of descriptors may still be used efficiently to discriminate among cultivars (Lahoz et al. 2011). The main differences emerging from several studies is that wild cardoon accessions collected either in central Italy (Noorani et al. 2012b), Sicily and Sardinia (Portis et al. 2005), or Spain (Lahoz et al. 2011) are genetically well separated from one another, probably due to a restricted gene flow despite the allogamous nature of cardoon. Wild cardoon accessions were also evaluated using molecular markers in order to assess the degree of variability between and within populations. The results provide valuable information on the genetic diversity present in wild cardoon populations in Italy. The populations were well segregated and, in contrast to artichoke, there were correlations between the genetic and the geographic distances. Nevertheless, with some exceptions, (for example a genotype from Tarquinia appeared to be closer to a Siena genotype (Noorani et al. 2012b)) indicate a high levels of diversity present in Tarquinia. In addition, the diversity levels found within each population were higher than expected, a positive sign with regard to the conservation of local genetic diversity.

The populations of wild cardoons studied are located in small areas and therefore at risk of bottleneck effects, inbreeding depression and possible extinction. Actions are currently underway to classify the area around Tarquinia (called Monti della Tolfa) as a regional park. This is an important step for the preservation of these wild populations as the distribution and extent of diversity present in the wild in Italy

is not certain. While there might be an intrinsic value to conserving diversity, the practical aspects here are that the wild cardoon is of great interest for incorporating possible disease and pest resistance genes in artichoke production (Cirulli et al. 1994; Ciccicarese et al. 2012), as a source of inulin (Melilli and Raccuia 2012), oil extract (Raccuia et al. 2011, 2012a), or genes for tolerance to salt stress (Raccuia et al. 2004). The cultivated cardoon collections have a remarkable level of genetic variation both between and within accession (Migliaro et al. 2012; Raccuia et al. 2012b). Studies of wild cardoon populations collected in different areas of Sicily exhibited variation for abiotic stresses, such as salinity and water stress resistance during seed germination (Raccuia et al. 2004). A correlation was also found between genetic variation and geographical origin among seven populations of wild cardoon from Sicily and Sardinia (Portis et al. 2005). Raccuia et al. (2004) found similar correlation for Sicilian wild cardoon populations.

When comparing accessions of different *Cynara cardunculus* species, based on the percentage of shared molecular markers' alleles and Nei's genetic distance, it emerges that wild cardoons from the Eastern Mediterranean were closer to artichoke than to cultivated cardoon (Sonnante et al. 2008), in comparison to wild cardoons from the Western Mediterranean, and that the genetic distance between the two wild cardoon gene pools was high.

23.3.4 DNA Conservation

CYNARES project partners extracted DNA from all accessions studies, and most of it has been stored in the DNA bank held at the Institute of Plant Genetics, Bari (Italy). Molecular analyses were carried out using AFLP, ISSR and microsatellite markers. Discriminant analysis classified 98 % of the analysed genotypes with accessions clustering based on their genetic distance; moreover, the analyses detected the correct assignment of each genotype with respect to country of origin and artichoke typologies. The dendrogram showing Nei's genetic distance between all the analysed accessions was able to cluster together all the accessions with cardoon germplasm, in agreement with the results of Cravero et al. (2010), Sonnante et al. (2008) and Mauro et al. (2009). On the other hand, accessions with same names but obtained by different Institutions such as Camerys, Puvisameliore, Salambo, Salanquet, and Violet du Gapeau or the different selections derived from Violet Provence were located far apart in the dendrogram. This highlights the problem of inaccurate cultivar definition over years of cultivation in various locations. As previously reported, cultivars are often named according to their place of cultivation, but not necessarily because they are of different material. Conversely, differing accessions may be given similar names due to geographic area of cultivation (Boury et al. 2012).

Some accessions have private alleles (i.e. fragments present only in that accession); potentially, these alleles could be used to identify accessions by genetic fingerprinting. It is interesting to note that the accessions from the Salanquet and

Salambo (France) collection have a private band not present in the same accession-name coming from the Catania collection. The number of private bands found here is comparable to that observed in different germplasm collections of globe artichoke.

The different markers give slightly different pictures: the polymorphism information content (PIC) was on average higher for AFLPs markers (about 16 %) than for ISSR markers (about 5 %) indicating the former as a more informative marker. The values were lower than the one found by other authors. The Nei genetic diversity measured as $G^{st} = (HT-Hep)/HT$, with Hep indicating the average diversity within populations and HT indicating the total diversity, was, for the overall core collection, equal to 0.42 and with a great proportion of variation within accessions than between accessions. On average about 68 % of the variation could be attributed to differences within accessions. Overall, the percentage of polymorphism in the CYNARES collection is only about 19 %. Again, the different markers used detected different levels of polymorphism; in particular ISSR markers detected lower levels of variation within populations and a higher variation among populations.

23.4 Conclusions

With a reservoir of traits and characteristics, crop genetic variation overall allows varietal diversification, diversity in foods and farming methods, and the provision of materials for plant breeding. Focusing on these aspects, the seven CYNARES project partners collected and organized collections to create a unique database with morphological, biochemical and molecular characterization over the four-year project duration (Pagnotta 2011a; Pagnotta et al. 2012). Most accessions were directly collected from farmers' fields, in Italy, Spain and France (the list of accessions are present in the CYNARES database at www.cynares.com and published in a booklet (Pagnotta 2011b)).

Local crop diversity, including landraces, is a key resource towards developing resilience in agro-ecosystems. Local varieties are better suited to local ecosystems, climatic conditions and farming practices. The importance of landraces is therefore two-fold: they are of direct use, particularly to smallholder farmers, and they constitute a potential source of basic genetic material for developing better adapted improved varieties in the future (FAO 2010). In this context, *in vitro* collections of artichoke landraces have been established, including: (a) 25 clones of Romanesco type, (b) four landraces from the Marche region, (Monte Lupone A, Monte Lupone B, Jesino, Ascolano), (c) one landrace named Pisa from the Tuscany region, (d) 2 landraces from the Campania region (Bianco di Pertosa, Tondo Rosso di Paestum).

In the past, efforts to counter genetic erosion of agricultural diversity have focused on *ex situ* conservation of germplasm (genebanks). The best strategies, however, combine *ex situ* conservation with *in situ* conservation, which includes on-farm management of varietal diversity. The need to address and integrate *in situ* and *ex situ* conservation has been recognized by international fora and agreements, including

the Convention on Biological Diversity (CBD), the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) and the recently adopted Second Global Plan of Action for PGRFA (Second GPA). The project integrated both approaches, as it combined *in situ* on-farm conservation and use along with the development of a DNA bank for *Cynara cardunculus*. The bank currently holds: 510 individuals belonging to 145 accessions of artichoke; 66 individuals belonging to 14 varieties of cultivated cardoon; 301 individuals belonging to 39 populations of wild cardoon.

The conservation of crop wild relatives has received much attention due to high rates of genetic erosion occurring in commonly cultivated crop species (Smale et al. 2004; Hou and Gao 1999). These wild species are at even greater risk of genetic erosion due to the destruction of natural habitats, and to globalization which causes a reduction in the diversity of species and planted varieties. It is therefore essential to more accurately estimate existing diversity to be able to efficiently conserve and manage these resources.

References

- Acquadro A, Portis E, Albertini E, Lanteri S (2005) M-AFLP-based protocol for microsatellite loci isolation in *Cynara cardunculus* L. (*Asteraceae*). *Mol Ecol Notes* 5:272–274
- Acquadro A, Lanteri S, Scaglione D et al (2009) Genetic mapping and annotation of genomic microsatellites isolated from globe artichoke. *Theor Appl Genet* 118:1573–1587
- Acquadro A, Papanice MA, Lanteri S et al (2010) Production and fingerprinting of virus-free clones in a reflowering globe artichoke. *Plant Cell Tiss Organ Cult* 100:329–337
- Agostinucci G, Loseby M (2007) Organizzazione e competitività—i problemi del settore. *Inf Agrario (Speciale carciofo)* 22:35–38
- Ancora G, Crinò P, Tavazza R et al (2012) The first three clones selected from the traditional artichoke ‘Romanesco’ populations and proposed for the release of new varieties. *Acta Hort* 942:125–131
- Angelini LG, Ceccarini L, Nasso Di Nasso N, Bonari E (2009) Long-term evaluation of biomass production and quality of two cardoon (*Cynara cardunculus* L.) cultivars for energy. *Biomass Bioenerg* 33:810–816
- Antunes A, Amaral E, Belgacem MN (2000) *Cynara cardunculus* L.: chemical composition and soda-antraquinone cooking. *Ind Crops Prod* 12:85–91
- Basnizki J, Zohary D (1994) Breeding of seed-planted artichoke. *Plant Breed Reviews* 12:253–269
- Bianco VV (1990) Carciofo (*Cynara scolymus* L.). In: Bianco VV, Pimpini F (eds) *Orticoltura*. Patron Editore, Bologna (in Italian), pp 209–251
- Bianco VV (2005) Present situation and future potential of artichoke in the Mediterranean basin. *Acta Hort* 681:39–55
- Bonasia A, Conversa G, Lazziaera C et al (2010) Morphological and qualitative characterization of globe artichoke head from new seed propagated cultivars. *J Sci Food Agric* 90:2689–2693
- Boury S, Jacob A-ME, Egea-Gilabert C et al (2012) Assessment of genetic variation in an artichoke European collection by means of molecular markers. *Acta Hort* 942:81–87
- Brown AHD (1989a) The case for core collections. In: Brown AHD, Frankel DH, Marshall DR, Williams JT (eds) *The use of plant genetic resources*. Cambridge University Press, Cambridge, pp 136–156
- Brown AHD (1989b) Core collections: a practical approach to genetic resources management. *Genome* 31:818–824

- Bruneton J (1995) Pharmacognosy phytochemistry medicinal plants. Lavoisier Secaucus N.Y. pp 218–219
- Cadinu M, Baghino L, Mallica G et al (2012) Collection of artichoke germplasm from different Mediterranean regions. *Acta Hort* 942:103–108
- Castiglione V, Cavallaro V, Di Silvestro I, Melilli MG (2007) Influence of different substrates on *in vitro* initiation of some early and late cultivars of globe artichoke (*Cynara cardunculus* L. Subsp. *Scolymus* (L.) Hayek). *Acta Hort* 730:107–112
- Ciancolini A, Alignan M, Miquel J et al (2013) Morphological characterization, biomass and pharmaceutical compound production from Italian globe artichoke genotypes. *Industrial Crops and Products* 49:326–333
- Ciancolini A, Rey N, Pagnotta MA, Crinò P (2012) Characterization of Italian spring globe artichoke germplasm: morphological and molecular profiles. *Euphy* 186(2):433–443
- Ciccarese F, Crinò P, Raccuia SA, Temperini A (2012) Use of resistant cardoons as rootstocks for the control of *Verticillium* wilt in globe artichoke. *Acta Hort* 942:201–205
- Cirulli M, Ciccarese F, Amenduni M (1994) Evaluation of Italian clones of artichoke for resistance to *Verticillium dahliae*. *Plant Dis* 78:680–682
- Cravero V, Martin E, Lopez Anido F, Cointy E (2010) Stability through years in a non-balanced trial of globe artichoke varietal types. *Sci Hort* 126:73–79
- Crinò P, Ciancolini A, Saccardo F et al (2011) Accessions of spring globe artichoke and cardoon conserved and evaluated by ENEA and the University of Tuscia. In: Pagnotta MA (ed) Evaluation of the European cynara germplasm extract from the cynares database. The University of Tuscia press. Via S.C. de Lellis Viterbo, Italy ISBN 978–88-87173–11–6
- Crinò P, Tavazza R, Rey Munoz NA et al (2008) Recovery, morphological and molecular characterization of globe artichoke ‘Romanesco’ landraces. *Gen Res Crop Evol* 55:823–833
- Curt MD, Sánchez G, Fernàndez J (2002) The potential of *Cynara cardunculus* L. for seed oil production in a perennial cultivation system. *Biomass Bioenergy* 23:33–46
- Egea-Gilabert C, Niirola D, Gómez P et al (2011) Accessions of cardoon and globe artichoke evaluated by Universidad Politécnica de Cartagena. In: Pagnotta MA (ed) Evaluation of the European cynara germplasm extract from the cynares database. The University of Tuscia press. Via S.C. de Lellis Viterbo, Italy ISBN 978–88-87173–11–6
- Engelmann F (1991) *In vitro* conservation of tropical plant germplasm. *Euphytica* 57:227–243
- FAO (2010) The second report on the state of the world’s plant genetic resources for food and agriculture, Rome
- FAOSTAT (2009) <http://faostat.fao.org/>
- Fernández J, Curt MD, Aguado PL (2006) Industrial applications of *Cynara cardunculus* L, for energy and other uses. *Ind Crops Prod* 24:222–229
- Foti S, Mauromicale G, Raccuia SA et al (1999) Possible alternative utilization of *Cynara* spp. Biomass, grain yield and chemical composition of grain. *Ind Crops Prod* 10:219–228
- Foury C (1989) Ressources genetiques et diversification de l’artichaut (*Cynara scolymus* L.). *Acta Hort* 242:155–166
- Foury C, Martin F, Pécaut P et al (2005) Avantages et difficultés de la création d’ hybrides F₁ d’artichaut à semer. *Acta Hort* 681:315–322
- Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber WK et al (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170
- Fratianni F, Tucci M, De Palma M et al (2007) Polyphenolic composition in different parts of some cultivars of globe artichoke (*Cynara cardunculus* L. var. *scolymus* (L.) Fiori). *Food Chem* 104:1282–1286
- Galán E, Prados F, Pino A et al (2008) Influence of different amounts of vegetable coagulant from cardoon *Cynara cardunculus* and calf rennet on the proteolysis and sensory characteristics of cheeses made with sheep milk. *Inter Dairy J* 18(1):93–98
- Gebhardt R (2002) Inhibition of cholesterol biosynthesis in HepG2 cells by artichoke extracts is reinforced by glucosidase pretreatment. *Phytoth Res* 16(4):368–372

- Gominho J, Fernandez J, Pereira H (2001) *Cynara cardunculus* L., a new fibre crop for pulp and paper production. *Ind Crops Prod* 13:1–10
- Gominho J, Laureço A, Curt M et al (2009) Characterization of hairs and pappi from *Cynara cardunculus capitula* and their suitability for paper production. *Ind Crops Prod* 29:116–125
- Gominho J, Lourenço A, Palma P et al (2011) Large-scale cultivation of *Cynara cardunculus* L., for biomass production—A case of study. *Ind Crops Prod* 33:1–6
- Hernández-Allica J, Becerril JM, Garbisu C (2008) Assessment of the phytoextraction potential of high biomass crop plants. *Env Pol* 152:32–40
- Hou XY and Gao WD (1999) The conservation and the application of the crops wild relatives. *Biol Diversity* 7(4):327–331
- Idrisi Z (2005) The Muslim agricultural revolution and its influence on Europe. Manchester, FSTC
- Ierna A, Mauromicale G (2010) *Cynara cardunculus* L. genotypes as a crop for energy purposes in a Mediterranean environment. *Biomass Bioenerg* 34:754–760
- Janeiro LV, Vieitez AM, Ballester A (1995) Cold storage of *in vitro* cultures of wild cherry, chestnut and oak. *Ann Sci For* 52:287–293
- Jouy C, Kermarrec D, Menard V et al (2011) Accessions of cardoon and globe artichoke varieties maintained and evaluated by GEVES, coordinator of the French CYNARA network. In: Pagnotta MA (ed) Evaluation of the European cynara germplasm extract from the cynares database. The University of Tuscia press. Via S.C. de Lellis Viterbo, Italy ISBN 978–88-87173–11–6
- Lahoz I, Fernández JA, Migliaro D et al (2011) Using molecular markers, nutritional traits and field performance data to characterize cultivated cardoon germplasm resources. *Sci Hort* 127:188–197
- Lanteri S, Portis E (2008) Globe artichoke and Cardoon. In: Prohens J, Nuez F (eds.) *Vegetables I: Asteraceae, Brassicaceae, Chenopodiaceae, and Cucurbitaceae*. Springer, pp 49–74
- Lanteri S, Di Leo I, Ledda L et al (2001) RAPD variation within and among populations of globe artichoke cultivar ‘Spinoso sardo’. *Plant Breeding* 120:243–246
- Lanteri S, Saba E, Cadinu M et al (2004a) Amplified fragment length polymorphism for genetic diversity assessment in globe artichoke. *Theor Appl Genet* 10:1534–1544
- Lanteri S, Acquadro A, Saba E, Portis E (2004b) Molecular fingerprinting and evaluation of genetic distances among selected clones of globe artichoke (*Cynara cardunculus* L. var. *scolymus* L.). *J Hort Sci Biotech* 79:863–870
- Lanteri S, Acquadro A, Comino C et al (2006) A first linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus* L.) based on AFLP, SSAP, M-AFLP and microsatellite markers. *Theor Appl Genet* 112:347–357
- Lanteri S, Portis E, Acquadro A et al (2012) Morphology and SSR fingerprinting of newly developed *Cynara cardunculus* genotypes exploitable as ornamentals. *Euphytica* 184:311–321
- Lattanzio V, Kroon PA, Linsalata V, Cardinali A (2009) Globe artichoke: a functional food and source of nutraceutical ingredients. *J Funct Foods* 1:131–144
- Llorach R, Espin JC, Tomás BFA, Ferreres F (2002) Artichoke (*Cynara scolymus* L.) By-products as a potential source of health-promoting antioxidant phenolics. *J Agric Food Chem* 50:3458–3464
- Lo Bianco C, Fernández JA, Migliaro D et al (2011) Identification of F1 hybrids of artichoke by ISSR markers and morphological analysis. *Mol Breeding* 27:157–170
- Lo Bianco C, Saccardo F, Olimpieri I et al (2012) Floral biology in male sterile clones of globe artichoke (*Cynara cardunculus* subsp. *Scolymus* (L) Hegi). *Acta Hort* 942:159–164
- Lombardo S, Pandino G, Mauromicale G et al (2010) Influence of genotype, harvest time and plant part on polyphenolic composition of globe artichoke (*Cynara cardunculus* L. var. *scolymus* (L.) Fiori). *Food Chem* 119:1175–1181
- Macua JJ, Lahoz I, Garnica J et al (2011) Accessions of cardoon and globe artichoke recovered and evaluated by Instituto Técnico y de Gestión Agrícola (ITGA) In: Pagnotta MA (ed) Evaluation of the European cynara germplasm extract from the cynares database. The University of Tuscia press. Via S.C. de Lellis Viterbo, Italy ISBN 978–88-87173–11–6
- Mallica G, Cadinu M, Repetto A (2004) Results of the clonal selection of artichoke cv Spinoso Sardo (*Cynara scolymus* L.; Sardinia). *Italus Hortus* 11:25–28

- Mauro R, Portis E, Acquadro A et al (2009) Genetic diversity of globe artichoke landraces from Sicilian small-holdings: implication for evolution and domestication of the species. *Conserv Genet* 10:431–440
- Mauromicale G, Ierna A (2000) Panorama varietale e miglioramento genetico del carciofo. *Informatore Agrario* 56:39–45
- Melilli MG, Raccuia SA (2012) Inulin and inulin metabolizing enzyme activities during the growth cycle of wild cardoon. *Acta Hort* 942:419–425
- Melilli MG, Tringali S, Riggi E, Raccuia SA (2007) Screening of genetic variability for some phenolic constituents of globe artichoke head. *Acta Hort* 730:85–91
- Menin B, Comino C, Moglia A et al (2010) Identification and mapping of genes related to caffeoylquinic acid synthesis in *Cynara cardunculus* L. *Plant Sci* 179:338–347
- Messmer M, Scheider E, Stekly G, Büter B (2002) Determination of progenitors and the genetic stability of the artichoke cultivar Saluschoke using molecular markers. *J Herbs Spices Med Plants* 9:177–182
- Migliaro D, Gómez diMP, Esteban A et al (2012) Genetic variability in ten Spanish cardoon populations as assessed by morphological, agronomical and molecular analyses. *Acta Hort* 942:115–122
- Mondini L, Noorani A, Pagnotta MA (2009) Assessing plant genetic diversity by molecular tools. *Diversity* 1:19–35
- Montemurro F, Sarli G, Montesano V et al (2012) Morpho-agronomic characterization of artichoke genotypes belonging to the institute of plant genetics (Igv-Cnr) collection. *Acta Hort* 942:109–114
- Muntoni M, Poddie M (2002) Clonal selection of “Spinoso Sardo” artichoke. Assessment of attitude to nursery growing in micropropagated plants (*Cynara cardunculus* var *scolymus* L.—Sardinia). *Italus Hortus* 9(3):67–69
- Noorani A, Crinò P, Rey N et al (2012a) Diversity assessment of seven Italian globe artichoke varieties using agromorphological parameters. *Acta Hort* 942:95–101
- Noorani A, Rey N, Temperini A et al (2012b) Assessment of genetic variation in three populations of Italian wild cardoon. *Acta Hort* 942:49–54
- Pagnotta MA (2011a) Genetic resources of cynara spp. an AGR GEN RES European Project CYNARES. *Kew Bull* 65:555–560
- Pagnotta MA (2011b) Evaluation of the European cynara germplasm extract from the cynares database. The University of Tuscia press. ISBN 978–88-87173–11–6
- Pagnotta MA, Cardarelli MT, Rey et al (2004) Assessment of genetic variation in artichoke of ‘Romanesco’ type by molecular markers. *Acta Hort* 660:99–104
- Pagnotta MA, Saccardo F, Temperini O et al (2012) Characterization of the *Cynara* European genetic resources. *Acta Hort* 942:89–93
- Pandino G, Fraser LCourtsetal (2010) Caffeoylquinic acids and flavonoids in the immature inflorescence of globe artichoke, wild cardoon, and cultivated cardoon. *J Agric Food Chem* 58:1026–1031
- Pandino G, Lombardo S, Mauromicale G (2011a) Chemical and morphological characteristics of new clones and commercial varieties of globe artichoke (*Cynara cardunculus* var. *scolymus*). *Plant Foods Hum Nutr* 66:291–297
- Pandino G, Lombardo S, Mauromicale G, Williamson G (2011b) Profile of polyphenols and phenolic acids in bracts and receptacles of globe artichoke (*Cynara cardunculus* var. *scolymus*), germplasm. *J Food Comp Anal* 24:148–153
- Pignone D, Sonnante G (2004) Wild artichokes of south Italy: did the story begin here? *Gen Res Crop E* vol 51:577–580
- Porceddu E, Dellacecca V, Blanco VV (1976) Classificazione numerica di cultivar di carciofo. *Proceedings II International Congress of Artichoke*, Minerva Medica, Turin, pp 1105–1119
- Portis E, Mauromicale G, Barchi L et al (2005) Population structure and genetic variation in autochthonous globe artichoke germplasm from Sicily Island. *Plant Sci* 168:1591–1598

- Portis E, Mauromicale G, Mauro R et al (2009) Construction of a reference molecular linkage map of globe artichoke (*Cynara cardunculus* var. *scolymus*). *Theor Appl Genet* 120:59–70
- Raccuia SA, Melilli MG (2004) *Cynara cardunculus* L., a potential source of inulin in the Mediterranean environment: screening of genetic variability. *Aust J Agric Res* 55:693–698
- Raccuia SA, Melilli MG (2007) Biomass and grain oil yields in *Cynara cardunculus* L. genotypes grown in a Mediterranean environment. *Field Crops Res* 101:187–197
- Raccuia SA, Melilli MG (2010) Seasonal dynamics of biomass, inulin, and water-soluble sugars in roots of *Cynara cardunculus* L. *Field Crops Res* 116:147–153
- Raccuia SA, Mainolfi A, Mandolino G, Melilli MG (2004) Genetic diversity in *Cynara cardunculus* revealed by AFLP markers: comparison between cultivars and wild types from Sicily. *Plant Breeding* 123:280–284
- Raccuia SA, Piscioneri I, Sharma N, Melilli MG (2011) Genetic variability in *Cynara cardunculus* L. domestic and wild types for grain oil production and fatty acids composition. *Biomass Bioenergy* 35:3167–3173
- Raccuia SA, Melilli MG, Piscioneri I, Sharma N (2012a) Evaluation of fatty acids composition in grain oil of cardoon (*Cynara cardunculus* L.). *Acta Hort* 942:463–468
- Raccuia SA, Gallo G, Melilli MG (2012b) Effect of plant density on biomass and grain yields in *Cynara cardunculus* var. *atilis* cultivated in Sicily. *Acta Hort* 942:303–308
- Rice N, Cordeiro G, Shepherd M et al (2006) DNA banks and their role in facilitating the application of genomics to plant germplasm. *Plant Gen Res* 4:64–70
- Robba L, Carine MA, Russell SJ, Raimondo FM (2005) The monophyly and evolution of *Cynara* L. (*Asteraceae*) sensu lato: evidence from the internal transcribed spacer region of nrDNA. *Plant System E vol* 253:53–64
- Rottenberg A, Zohary D (1996) The wild ancestry of the cultivated artichoke. *Genet Resour Crop Evol* 43:53–58
- Saccardo F (2009) Miglioramento genetico In: Calabrese N (ed) *Il Carciofo e il Cardo*. Bayer Crop Science, Bologna, pp 286–297
- Sarkar D, Naik PS (1998) Cryopreservation of shoot tips of tetraploid potato (*Solanum tuberosum* L.) clones by vitrification. *Ann Bot* 82:455–461
- Smale M, Bellon MR, Jarvis D, Sthapit B (2004) Economic concepts for designing policies to conserve crop genetic resources on farms. *Genet Resour Crop Evol* 51:121–135
- Snyder MJ (1979) Investigation of propagational techniques for artichoke. In: *Atti 3rd Congr Int Stud Carciof*, Bari Ind Grafica Laterza, Bari, pp 347–358
- Sonnante G, De Paolis A, Lattanzio V, Perrino P (2002) Genetic variation in wild and cultivated artichoke revealed by RAPD markers. *Genet Res Crop E vol* 49:247–252
- Sonnante G, De Paolis A, Pignone D (2004) Relationships among artichoke cultivars and some related wild taxa based on AFLP markers. *Plant Genet Resour* 1:125–133
- Sonnante G, Pignone D, Hammer K (2007) The domestication of artichoke and cardoon: from Roman times to the genomic age. *Annals Bot* 100:1095–1100
- Sonnante G, Carluccio AV, De Paolis A, Pignone D (2008) Identification of artichoke SSR markers: molecular variation and patterns of diversity in genetically cohesive taxa and wild allies. *Genet Resour Crop Evol* 55:1029–1046
- Sonnante G, D'Amore R, Blanco E et al (2010) Novel hydroxycinnamoyl-coenzyme A quinate transferase genes from artichoke are involved in the synthesis of chlorogenic acid. *Plant Physiol* 153:1–15
- Sonnante G, Sarli G, Di VD (2011a) Characterization of globe artichoke accessions belonging to the *Cynara* collection of the Institute of Plant Genetics, CNR, Italy. In: Pagnotta MA (ed) *Evaluation of the European cynara germplasm extract from the cynares database*. The University of Tuscia press. Via S.C. de Lellis Viterbo, Italy ISBN 978–88–87173–11–6
- Sonnante G, Gatto A, Morgese A et al (2011b) Genetic map of artichoke x wild cardoon: toward a consensus map for *Cynara cardunculus*. *Theor Appl Genet* 123:1215–1229
- Soressi GP (2003) Available variability usable for breeding of globe artichoke (*Cynara cardunculus* L. var. *scolymus* L.). *Inf Agrario* 59:47–50

- Sousa MJ, Malcata FX (1997) Comparison of plant and animal rennets in terms of microbiological, chemical and characteristics of ovine cheese. *J Agric Food Chem* 45:74–81
- Sousa MJ, Malcata FX (1998) Identification of peptides from ovine milk cheese manufactured with animal rennet or extracts of *Cynara cardunculus* as coagulant. *J Agric Food Chem* 46:4034–4041
- Tivang J, Skroch P, Nienhuis J, De Vos N (1996) Randomly amplified polymorphic DNA (RAPD) variation among and within artichoke (*Cynara scolymus* L.) cultivars and breeding populations. *J Am Hortic Sci* 121:783–788
- Withers LA, Wheelans SK, Williams JT (1990) *In vitro* conservation of crop germplasm and the IBPGR databases. *Euphytica* 45:9–22
- Zaniboni R (2009) Ibridi commerciali. AA.VV. Il carciofo e il cardo. Coordinamento scientifico N. Calabrese. Collana Coltura e cultura. Bayer Crop Science. Ed. Script, Bologna

Chapter 24

Analysis and Exploitation of Cereal Genomes with the Aid of *Brachypodium*

Hikmet Budak, Pilar Hernandez and Alan H. Schulman

Contents

24.1	Syteny and its Applications	586
24.2	<i>Brachypodium</i> as a Model for the Temperate Cereals	587
24.2.1	Why is a Model Useful?	587
24.2.2	<i>Brachypodium</i> as a Model Species	588
24.3	<i>Brachypodium</i> Genetics and Genomics	589
24.3.1	Phylogenetics	589
24.3.2	<i>Brachypodium</i> Mendelian Genetics	589
24.3.3	<i>Brachypodium</i> Cytogenetics	589
24.3.4	Syntenic Relationships	590
24.3.5	<i>Brachypodium</i> genomics	591
24.4	<i>Brachypodium</i> Resources	591
24.4.1	Germplasm Collections	591
24.4.2	<i>Brachypodium</i> Methodological Resources	594
24.4.3	Bioinformatic Resources	595
24.5	<i>Brachypodium</i> as an Aid for Cereal Genetics and Genomics	595
24.5.1	Polyploidization	595
24.5.2	Assembly of Sequencing Data	596
24.5.3	Markers for Cereal Gene Mapping	597
24.5.4	Map-based Cloning and Genetic Mapping	599
24.5.5	Comparative Studies Using Bioinformatics	599
24.6	Analysis in <i>Brachypodium</i> of Traits for Cereal Crops	602
24.6.1	Biotic Stress Resistance	602

A. H. Schulman (✉)

Department of Biotechnology and Food Research,
MTT Agrifood Research, 31600 Jokioinen, Finland

Institute of Biotechnology, Viikki Biocenter,
University of Helsinki, P.O. Box 65, 00014 Helsinki, Finland
e-mail: alan.schulman@helsinki.fi

H. Budak
Biological Sciences and Bioengineering Program,
Faculty of Engineering and Natural Sciences,
Sabanci University, Orhanli, Tuzla-Istanbul, Turkey
e-mail: budak@sabanciuniv.edu

P. Hernandez
Institute for Sustainable Agriculture (IAS-CSIC),
Alameda del Obispo s/n, 14080 Córdoba, Spain

24.6.2	Abiotic Stress Tolerance	603
24.6.3	Bioenergy	604
24.6.4	Yield, Grain Characteristics and Plant Development	605
24.7	Direct Agricultural Use of <i>Brachypodium</i>	606
24.8	Conclusions	606
	References	606

Abstract *Brachypodium* was proposed to become the *Arabidopsis* of the cereals due to its small stature, rapid life cycle, phylogenetic proximity to the “core Pooids,” and small genome. Due to the availability of a high quality genome sequence and the development of many tools for functional genomics, it has lived up to this promise. Here, the biology, genetics, and genomics of *Brachypodium* will be reviewed as a context for the use of the plant, particularly the annual diploid *B. distachyon*, as a research system. The available resources will be summarized. The use of *Brachypodium* as a tool for research on the cereal crops will be presented, as will current research in *Brachypodium* itself on traits relevant to grain and bioenergy production.

Keywords *Brachypodium* · Synteny · Model plants · Triticeae · Pooideae · Aveneae · Barley · Wheat · Phylogenetics · Cytogenetics · Genomics · Genome evolution · Germplasm collections · Mutagenesis · TILLING · T-DNA insertion lines · Transformation · Gene silencing · VIGS · Bioinformatics · Polyploidization · Molecular markers · Genome sequence · Map-based cloning · Transcription · Gene function · Biotic stress · Abiotic stress · Bioenergy · Yield · Plant architecture · Grain quality

24.1 Synteny and its Applications

The use of *Brachypodium* in the mining of cereal genes has, at its heart, the exploitation of synteny within the grasses. Restriction-fragment length polymorphisms (RFLPs) in cloned genes served as the first-generation DNA markers that were based on filter hybridization (Botstein et al. 1980). The RFLP markers, which were gene-based, could cross-hybridize to the target genes in related species. When Gale and coworkers developed RFLP maps in hexaploid bread wheat (*Triticum aestivum*) and its diploid ancestors *Triticum urartu* and *Aegilops squarrosa*, as well as in the crops barley (*Hordum vulgare*) and rye (*Secale cereale*), also from the Triticeae tribe, they noticed great similarities in the genetic maps. Alignment of the maps and extension of the analysis to other grasses led them to propose that 19 linkage blocks, ranging from chromosome segments to whole chromosomes, were all that was needed to reconstruct the ancestral cereal genome and its modern descendents (Moore et al. 1995).

The model that has developed, based on these early studies, is that grass chromosomes contain extensive blocks of genes that are maintained in conserved order (collinearity) and that these blocks are present on chromosomes related by descent (synteny) such as the homeologous chromosome sets in the Triticeae, which include the A, B, and D genomes of bread wheat and the H genome of barley (Bolot et al. 2009). The advent of large-scale genome sequencing made it possible to reconstruct both the fine-scale relationships of syntenic genes across related genomes and to model the paleogenomics of grass evolution, particularly the duplications, division, and fusion of chromosomes and their segments (Salse et al. 2009).

Chromosomes over evolutionary time have suffered the vagaries of fusions, fissions, deletions, recombinations, and duplications, which have affected both their number and linear integrity. Nevertheless, the preservation of gene order both on a large scale (macrocollinearity) and locally (microcollinearity) has proven to be very useful for evolutionary studies and gene isolation, as will be presented in this chapter. The overall conclusion is that while retrotransposons have replicated and expanded genome size and homeologous chromosomes, gene order has been largely preserved. The occurrence of collinearity and synteny is not limited to the grasses. It has been found so far to a lesser extent, for example, within the Solanaceae (Wu et al. 2009), Fabaceae (Ellwood et al. 2008), and Rosaceae (Cabrera et al. 2009). However, the discussion will be focused here on the member of syntenic “circle” of grasses with the smallest genome, *Brachypodium*.

24.2 *Brachypodium* as a Model for the Temperate Cereals

24.2.1 *Why is a Model Useful?*

Forage grasses and temperate cereals, including the tribes Triticeae and Aveneae of the Pooideae, are economically important for food and feed. Their domestication from wild grassland species took place between 3,000 and 10,000 years ago in several episodes (Brown et al. 2009). This process created population bottlenecks, resulting in a loss of genetic diversity in modern cultivated varieties compared to wild ancestors. For wheat, the remaining variation is estimated to be 10–20 % of the wild variation of its ancestor (Langridge et al. 2006). Securing the food and feed supply of the future will require crops with high yields under low input and challenges from disease and climate (Rosegrant and Cline 2003). Genomics will provide a key to the genes needed, in combination, to provide such crops.

The idea of using a model rather than the target crop itself is to overcome one or more limitations. Generally a model offers a rapid life cycle, collections of mutants and gene-tagged lines, easy genetics, easy and compact cultivation, genome sequence, large collaborative scientific community, and sufficient biological similarity to the target crop for transfer of information from the model to the target. With the sequencing of crop genomes and the development of tool sets for them, the

line between model and target has begun to blur. For example, a diploid Triticeae cereal may become a model for a tetraploid or hexaploid cereal, and a sequenced hexaploid a model for an unsequenced one, as the genomic tools improve. Nevertheless, *Brachypodium* offers sufficient interest as a widespread wild pooid with well-developed tools that its usefulness to biology in it should withstand progress in the sequencing of complex grass genomes.

24.2.2 *Brachypodium as a Model Species*

Given the synteny and collinearity among the grass genomes described above, it would appear that any member of the “crop circle” of grasses can be an entry point for exploitation of the genomic relationships for crop improvement. While this is generally true, the large and complex genomes of forage grasses and some species within the Triticeae are barriers in genomics research and molecular breeding. *Brachypodium distachyon* was first proposed as a model grass for functional genomics already in 2001 (Draper et al. 2001) for a combination of reasons. Phylogenetically, *Brachypodium* belongs to the Brachypodieae tribe, which diverged from the Pooideae subfamily just prior to the radiation of modern “core pooids.” These pooids include the tribes Triticeae, Bromeae, Poeae, and Avenae, whose species include barley, oat, rye, and various fodder grasses. Hence, its phylogenetic position, having diverged just before the clade containing “core pooid” species including wheat, offers biological relevance to crop improvement. It has the smallest genome size in grasses identified so far. The species is a self-fertile annual, with a life cycle less than 4 months long, accelerating experiments. Its small stature of about 20 cm, combined with non-shattering spikes and easy cultivation, makes it amenable to growth in high densities and numbers for genetics. Tissue culture and transformation protocols offer opportunities for reverse genetics. These features will be examined in greater detail below.

A single species cannot serve as a model organism for all species for all biological traits. Arabidopsis, rice (*Oryza sativa*), ryegrass (*Lolium*), and barley (*H. vulgare*) can be used as model organisms with their own advantages and disadvantages. Ryegrass and barley are easy to grow and have great value as a forage grass and a cereal respectively. Additionally, ryegrass is physically small. Barley also has important functional genomics resources including TILLING populations (Talame et al. 2008). *Brachypodium* is valuable with its intermediate position in evolution between the core Pooideae and the Ehrhartoideae. Although *Brachypodium* cannot be introgressed easily into other cereals and lacks agricultural value itself, it can be accepted as a low-cost, high-gain preferred model organism because of its properties appropriate to stress, grain yield and biofuel research. The present infrastructure for *Brachypodium* as will be presented in Sect. 4.

24.3 *Brachypodium* Genetics and Genomics

24.3.1 *Phylogenetics*

The *Brachypodium* genus in the Brachypodieae tribe contains 15–18 species including *Brachypodium distachyon* (purple false brome). Its natural range is the Mediterranean basin, southwest Asia, the Middle East and northeast Africa. With recent colonization, it has also been widely naturalized in Australia, America, South Africa and UK (Schippmann 1991; Catalán 2003; Stace 2010).

Studies to establish the phylogenetic relationships among Brachypodieae have shown that perennial species with long rhizomes (*B. arbuscula*, *B. retusum*, *B. rufepstre*, *B. phoenicoides*, *B. pinnatum*, and *B. sylvaticum*) are more closely related to each other than to *B. mexicanum*, having short rhizomes, or to the annual *B. distachyon* of primary interest experimentally (Catalán et al. 1997; Catalán and Olmstead 2000; Azhaguvel et al. 2009). The phylogenetic relationships among Brachypodieae and other cereals in the Poaceae family were also developed based on sequence variation as measured with mean synonymous substitution rates (Ks) between orthologous genes. It was estimated that *B. distachyon* diverged from the common wheat (*Triticum aestivum*) ancestor 32–39 million years ago (MYA), from the ancestor of cultivated rice (*Oryza sativa*) 40–53 MYA and from the branch leading to sorghum (*Sorghum bicolor*) 45–60 MYA (IBI 2010).

24.3.2 *Brachypodium Mendelian Genetics*

A genetic linkage map of *B. distachyon* was obtained with the genotyping-by-sequencing (GBS) approach in an F₂ mapping population. The method is essentially the sequencing of AFLP fragments, whereby both the fragment occurrence and the SNPs within the fragments are scored. An exceptionally high recombination rate was observed, higher in gene-rich regions and lower in repetitive regions and including centromeres. Moreover, positive correlation was detected between interspecific synteny and recombination rate (Huo et al. 2011).

24.3.3 *Brachypodium Cytogenetics*

The species of the Brachypodieae have various monoploid chromosome numbers ($x = 5, 7, 8, \text{ or } 9$). The sequenced species *B. distachyon* has a monoploid chromosome number of $x = 5$, with populations displaying either diploid, tetraploid, or hexaploid genomes ($2n = 10, 20 \text{ or } 30$). Recently, a variety of experiments were performed to investigate the evolution, origins and taxonomic split of three ploidy cytotypes. The cytotypes are distinct in their morphological and anatomical properties. The $2n = 10$

cytotypes are smaller and mostly require vernalization for flowering, whereas the $2n = 20$ and 30 cytotypes have large seeds and exhibit prominent anthesis (Schwartz et al. 2010).

Statistical analyses of phenotypic traits, as well as cytogenetic analyses estimating genome size with flow cytometry, fluorescent *in situ* hybridization (FISH), comparative chromosome painting (CCP), and phylogenetic analyses was carried out. These were combined with estimates of divergence times and evolutionary rates based on plastid (*ndhF*, *trnLF*) and nuclear (ITS, ETS, CAL, DGAT, GI) genes. The studies showed that $2n = 10$ and $2n = 20$ emerged from different lineages, which were subject to different mutation rates, whereas $2n = 30$ is derived from a hybridization between these two. Based on this evidence, the three cytotypes can be considered different species, and are referred to as *B. distachyon*, *B. stacei*, and *B. hybridum*, respectively (Catalán et al. 2012).

24.3.4 Syntenic Relationships

The high degree of synteny and orthology of *Brachypodium* with different members of the Poaceae family makes it a good structural model for the assembly of large genomes. Additionally, for most of the genes, it can serve as a good functional model. However, for positional gene cloning, comparative genomics, and genome assembly in the Triticeae, the multiple available genomes should preferably be used, because the syntenic relationships between the target species and any one particular region may be better with rice or another genome than with *Brachypodium*. Selective pressure may have led to divergence in the orthologous genes in one or other sequenced genome compared to the Triticeae ortholog (Yu et al. 2009; IBI 2010). In any case, there have been several genomic rearrangements in the evolution of wheat and barley after the divergence of *Brachypodium*. Nevertheless, because the evolutionary divergence of Ehrhartoideae (rice), Panicoideae (*Sorghum*), and Pooideae (*B. distachyon*, *T. aestivum*, and *Hordeum vulgare*) was relatively recent, the majority of genes and gene families are highly conserved between all of these temperate cereals.

Several studies on the orthology of individual *Brachypodium* and Triticeae genes have been made. These include the glutenin gene, the earliness per se *Eps-A* locus containing the *Mot1* and *FtsH4* genes, and stem rust resistance genes *Rpg1* and *Rpg4* (Drader and Kleinhofs 2010; Faricelli et al. 2010; Gu et al. 2010). When the orders of genes in large gene families were examined for different cereal genomes, most showed a high degree of conservation (IBI 2010). However, gene order in the nucleotide binding site (NBS), leucine-rich repeat (LRR) and F box gene families was shown not to be conserved (IBI 2010). In another report, the conserved ortholog of Hessian fly resistance gene H26 (previously mapped to its location in wheat 3DL) could not be identified in *Brachypodium*. Losses of conservation can result from rapid diversification under strong natural selection driven by pathogen pressure, in the case of NBS-LRR and the fly resistance gene, and by regulation of developmental and stress responses for F-boxes (Meyers et al. 2003; Xu et al. 2009).

24.3.5 *Brachypodium genomics*

B. distachyon was suggested as a model species for temperate cereals and forage grasses over a decade ago. In a short time, several genomic resources were rapidly established. Large expressed sequence tag (EST) libraries and databases were created. Highly refined cytogenetic markers were developed. In 2010, a very high quality genome sequence of accession Bd21 was published (IBI 2010). The final genome assembly is very complete, predicted to include 99.6 % of all the sequences based on paired-end information. In the initial annotation, a large number of ESTs were used; the annotation is of high quality (Vogel et al. 2006). Resequencing of *Brachypodium* accessions other than Bd21 is underway. Genomic markers for genetic screens are available. T-DNA tagged and EMS/fast neutron-mutated populations have been developed. *Brachypodium* expression and TILLING Affymetrix oligo-microarrays have been created. Additionally, several *Brachypodium* bioinformatic resources are available. *Brachypodium* bacterial artificial chromosome (BAC) libraries and a physical map based on these contigs are also present (Hasterok et al. 2004; Farrar and Donnison 2007; Gu et al. 2009; Huo et al. 2008, 2009).

An interesting area where *Brachypodium* will shed light on the Triticeae is the gain and loss of repetitive DNA, particularly of the retrotransposons. The retrotransposons in *B. distachyon* comprise 21.4 % of the genome, compared to 26 % in rice, and over 80 % in the Triticeae (IBI 2010). The numbers are sparse despite the recent activity of many elements in the genome, with 13 families younger than 20,000 years and 53 families less than 0.1 million years old. The genome appears, however, also to lose retrotransposons rapidly. The two long terminal repeats (LTRs) of many elements have recombined, leaving a solo LTR behind. In this way, an estimated 17.4 Mb of retrotransposon elements has been lost. In contrast, retroelements persist for very long periods of time in the Triticeae (Wicker and Keller 2007).

For the DNA transposons, Buchmann and colleagues compared 1 Mb of orthologous genomic sequences from *B. distachyon* and *B. sylvaticum*. They found that while a high percent of the genes in the region were collinear, only a low percentage of transposable elements were. They proposed a model in which double-strand break (DSB) repair causes insertions and deletions, transposons being a major factor in the erosion of intergenic sequences (Buchmann et al. 2012). For the third major repetitive component, a genome wide analysis of microsatellite distribution in different grasses including *B. distachyon* has been performed, which can later aid in microsatellite evolution studies in monocots and dicots (Sonah et al. 2011).

24.4 *Brachypodium* Resources

24.4.1 *Germplasm Collections*

24.4.1.1 Wild and Inbred Lines

Brachypodium germplasm collections have been assembled with the aim to include wide variation, uniform lines, and economically important traits (Filiz et al. 2009).

Until recently, *B. distachyon* germplasm collections were limited. Initial small collections included the USDA inbred lines (<http://www.ars-grin.gov/npgs>) Bd1-1, Bd2-3, Bd3-1, Bd18-1, Bd21, Bd21-3, and Bd29, as well as the Stace and Catalán collection from Spain, ABR1, ABR2, ABR3, ABR4, ABR5, ABR6, and ABR7. The need for extensive germplasm collections was pointed out 5 years ago (Garvin et al. 2008). At that point two main collections of *Brachypodium* were available: The USDA National Plant Germplasm System (NPGS, <http://www.ars-grin.gov/npgs/>; verified 7 June 2012) and the IBERS in Aberystwyth (Mur et al. 2011). Inbred lines have been developed from these collections, and designated with the prefix ‘Bd’ (Vogel et al. 2006; Garvin et al. 2008).

Currently, the largest one available is the Turkish collection established by Vogel et al. (2009) comprising 187 diploid lines from 53 locations and 84 inbreds. It harbors a high degree of variation is available (Filiz et al. 2009; Vogel et al. 2009). This collection is being expanded (Tuna et al. 2011). A large collection from Israel has been set up (Distelfeld et al. 2011) while population sampling has also been reported in Tunisia (Neji et al. 2011). Some of these collections are the basis for selection and inbred development programs in the corresponding institutes. Recombinant inbred lines were derived using a specific protocol, because initially plants failed to outcross owing to near cleistogamy (Routledge et al. 2004; Garvin et al. 2008).

There are several germplasm collections from Spain at the INIA (Soler et al. 2004, Hammami et al. 2011), University of Jaen (Manzaneda et al. 2012), UPM (Giraldo et al. 2012) and at IAS-CSIC (Pérez-Jiménez et al. 2009). Another large collection from Spain was used to develop inbred lines from various environments. Using this collection, sympatric $2n = 10$ and $2n = 30$ populations were detected in one location. Additionally, this collection was used to study intra-population and inter-population genotypic diversity in relation to adaptation. Inter-population comparisons were performed with Turkish lines and Bd21 (Vogel et al. 2009). A greater genetic diversity was observed in individuals from the west Mediterranean compared with those from the east, which was found to be the case for other temperate grasses such as *Hordeum marinum* (Jakob et al. 2007). While sympatric $2n = 10$ and $2n = 30$ populations have been detected in the available Spanish collections, *B. stacei* was only known from the type locality (Spain: Balearic Islands: Formentera) until recently. However, new studies have detected the presence of this species in other locations in southeastern and southern Spain (Hammami et al. 2011; Giraldo et al. 2012) and at the Spanish Canary Islands (Giraldo et al. 2012). The species could also be distributed in other Mediterranean localities (P. Catalán, personal communication).

24.4.1.2 EMS and Fast Neutron Populations for Genetic Screens

Mutant collections have been created using ethyl methanesulphonate (EMS) treatment (<http://brachypodium.pw.usda.gov/>) and fast neutron irradiation. In addition to their use in forward screens, these collections are also important in gene function identification by reverse genetics. TILLING (Targeted Induced Local Lesions in Genomes) is an approach used to identify functions of particular genes using an

EMS population (McCallum et al. 2000). A TILLING *B. distachyon* population was developed by INRA (http://www-ijpb.versailles.inra.fr/en/crb/crb_accueil.htm). The UTILLdb (<http://urgv.evry.inra.fr/UTILLdb>) is a commercial TILLING platform that includes *B. distachyon* lines and contains descriptions of phenotypes and information related to “tilled” genes related information (Dalmais et al. 2008).

24.4.1.3 T-DNA Collections and Insertional Mutagenesis

Using *B. distachyon* transformation techniques, efforts have been made to generate a T-DNA mutant library collections having known flanking sequences. The process includes generation of T-DNA mutant lines and analysis of the accessions to assign the flanking sequence tags (FSTs) to unique locations in the *B. distachyon* sequence. Thousands of lines have been established in USDA-ARS Western Regional Research Center (<http://Brachypodium.pw.usda.gov/TDNA>) and John Innes Centre (International *Brachypodium* Tagging Consortium BrachyTAG programme; <http://www.brachytag.org>). The Bd21T-DNA mutant plant lines in the context of the BrachyTAG programme (BrachyTAG.org) were produced with *Agrobacterium*-mediated transformation techniques (Thole et al. 2010; Thole and Vain 2012). The first technique developed for the retrieval of FSTs was adaptor ligation PCR coupled with sequencing (Vain et al. 2008; Thole et al. 2009). Genome walking has been one of the preferred choices for FST identification in T-DNA mutant populations and recently has been further improved and applied to *B. distachyon* (Taheri et al. 2012). New methods such as SiteFinding-PCR and its modified versions are being developed for isolating FSTs (Tan et al. 2005; Wang et al. 2011) and can be applied in *B. distachyon* mutant lines.

The use of T-DNA insertion lines has been demonstrated using the BrachyTAG collection in the studies of eukaryotic initiation factor 4A (eIF4A), brassinosteroid insensitive-1 (*BRI1*) and growth related genes (Thole et al. 2012). In a recent study, the function of eILF4A in stem elongation was examined with homozygous and hemizygous mutant plants in the BrachyTAG mutant population and by complementation studies of transforming *Brachypodium* with Arabidopsis eIF4A (Vain et al. 2011). In addition to gene trapping, the system is also useful for analysis of promoters and enhancers with gene-trapping and enhancer-trapping methods.

24.4.1.4 Mutant and Mapping Populations

A genetic linkage map of *Brachypodium* was obtained by genotyping by sequencing of single nucleotide polymorphisms (SNPs) using an F₂ mapping population. *Brachypodium* was observed to have a high recombination rate, higher in gene-rich regions and lower in repetitive regions and including centromeres. A positive correlation was detected between interspecific synteny and recombination rate (Huo et al. 2011).

24.4.2 *Brachypodium Methodological Resources*

24.4.2.1 Transformation and Regeneration Protocols

B. distachyon is one of most easily transformed grasses, which makes it a powerful functional genomics model. Both polyploid and diploid lines were transformed by bombardment (Draper et al. 2001; Christiansen et al. 2005). However the most popular *B. distachyon* transformation method is via *Agrobacterium*. In this regard, it is important to apply protocols (<http://Brachypodium.pw.usda.gov/>) to achieve high efficiency of transformation and low numbers of copies integrated. *Agrobacterium*-mediated transformation can be used both on diploid and polyploid *B. distachyon* lines. It is performed on compact embryogenic calli from immature embryos. Screening is performed via chemicals and in some cases coupled with phenotyping of transformed tissues and plants (Pacurar et al. 2008; Vain et al. 2008; Alves et al. 2009). *Agrobacterium* transformation has been useful in the production of Bd21 T-DNA mutants of the BrachyTAG programme (BrachyTAG.org; Vogel and Hill 2008; Thole et al. 2010, 2012).

24.4.2.2 Gene Silencing and VIGS

Virus-Induced Gene Silencing (VIGS) is a strategy to disrupt the expression of targeted genes. Barley stripe mosaic virus (BSMV), generally the silencing vector, is a single-stranded tripartite RNA virus. Infection and suppression of gene expression can be achieved via the rubbing of a recombinant virus genome including fragments of the gene of interest into the leaves of the host plant. This strategy has been used to suppress gene expression in barley and wheat (Holzberg et al. 2002; Scofield et al. 2005). It has been also been used to successfully silence a phytoene desaturase gene and genes involved in phosphate (Pi) uptake, specifically IPS1, PHR1, and PHO2 in *Brachypodium* (Demircan and Akkaya 2010; Pacak et al. 2010). Recent research has enhanced the BSMV VIGS system through the incorporation of an *Agrobacterium* delivery system and its coupling with a ligation-independent cloning (LIC) strategy for efficient cloning. These vectors were shown to down regulate-phytoene desaturase (PDS), magnesium chelatase subunit H (ChlH), and plastid transketolase (TK) gene expression in *B. distachyon* (Yuan et al. 2011).

24.4.2.3 Expression and TILLING Microarrays

An Affymetrix expression array was developed using the *B. distachyon* Bd21 genome sequence and ESTs. Unique single copy oligonucleotides were used to enable studies on gene-specific expression. Using this array a Bd21 expression atlas is being generated, which shows circadian clock, development, and stress related expression of *Brachypodium* genes. Recently, a miRNA microarray study was performed to identify drought responsive *Brachypodium* miRNAs in leaf and root tissues (Budak and Akpinar 2011)

24.4.3 Bioinformatic Resources

Genomic and other *Brachypodium*-related data are being collected and curated. In one website (<http://www.Brachypodium.org>), there is access to the *B. distachyon* 8 × assembly of the genome, ESTs and Affymetrix array probes. The website also contains a BLAST tool. Another website (http://www.gramene.org/Brachypodium_distachyon) includes *B. distachyon*-related information and a tool for genomic comparisons among species. The Gramene database (<http://www.gramene.org>) is currently a key resource for model and crop plants including *Brachypodium*. It contains several lines of information including quantitative trait loci (QTL), metabolic pathways, genetic diversity, genes, proteins, germplasm resources, literature, ontologies, markers, sequences, and maps from various studies (genetic, physical, bin). It also contains web services, including an Ensembl genome browser, a distributed annotation server (DAS), BLAST and a public MySQL (Youens-Clark et al. 2011).

The *B. distachyon* physical map can be accessed and compared with rice and sorghum genomes using another website (<http://www.modelcrop.org/>). Elsewhere (<http://www.phytozome.net>), orthologous or homologous genes in different plants can be found by sequence comparisons and phylogenetic relationships can be deduced. Genome-wide SSR markers can be downloaded from the BraMi (*Brachypodium* microsatellite markers) database (Sonah et al. 2011). Recently, a database called GramineaeTFDB, listing putative crop transcription factors (TFs), has been generated. *Brachypodium distachyon* TFs can be accessed from the website with their related information including sequence, promoter and domain features, assigned gene ontologies, and FL-cDNA information (Mochida et al. 2011b). A database (<http://markers.btk.fi>) for listing predicted molecular markers is also available (Rudd et al. 2005), as is a database (<http://phymap.ucdavis.edu/Brachypodium>) allowing access to the *B. distachyon* physical map (Gu et al. 2009). Recently, a database of chloroplast genome SSRs was created (Melotto-Passarin et al. 2011). Using *in silico* methods, *B. distachyon* microRNAs were identified and can be accessed from MIRBASE (Unver and Budak 2009).

24.5 *Brachypodium* as an Aid for Cereal Genetics and Genomics

24.5.1 Polyploidization

Several cereals including wheat are allopolyploids. *Brachypodium* lines were accepted to be autopolyploids ($2n = 20$; $2n = 30$), but recent research has shown that there are also autopolyploid races of *Brachypodium* including *B. retusum* ($2n = 38$), *B. pinnatum* ($2n = 28$), and *B. phoenicoides* ($2n = 28$) (Roberts et al. 1981; Wolny and Hasterok 2009). These studies on the phylogeny and evolution of *Brachypodium* chromosomes were conducted with *in situ* hybridization experiments.

Recently, the effect of old and new polyploidization events on the organization and function of the bread wheat genome was studied using wheat RNA sequencing data. *B. distachyon* was used as the reference to classify genes as either orthologous, paralogous, or homoeologous and for modeling changes in the grain genes in response to evolutionary events such as duplication, polyploidization, and speciation. The evolutionary times necessary for a given amount of functional and structural gene loss were estimated (Pont et al. 2011). The complex evolution of *Brachypodium* makes it an appropriate functional genomics model for the studies on the mechanism of polyploidization and genes involved in the process (Ozdemir et al. 2008). The *Ph1* locus, which plays a role in the diploidization of allohexaploid wheat was mapped using markers from a smaller orthologous region in *B. sylvaticum* (Griffiths et al. 2006).

24.5.2 Assembly of Sequencing Data

The *B. distachyon* genome has been structurally characterized and its synteny with wheat and rice was assessed. Annotated *B. distachyon* genes were BLAST searched against the wheat EST database and wheat ESTs mapped to deletion bins. The work suggested that *B. distachyon* will aid in ordering wheat ESTs and developing markers for targeted wheat genomic regions because some *B. distachyon* BACs gave hits to multiple ESTs mapped to the same deletion bins (Huo et al. 2009). Two BAC libraries were constructed using the inbred diploid line Bd 21, representing 19.2-fold of combined genome coverage. BAC-end sequences (BESs) were blasted against NCBI GenBank and GIRI repeat databases, suggesting that a considerable proportion of the *B. distachyon* genome was transcribed, but a low proportion was formed of repeats.

The closer relationship of *Brachypodium* to the Triticeae rather than other grasses was once more shown after a blast of BESs to wheat and maize EST databases. Those having significant matches to wheat ESTs were mapped to individual chromosome bin positions. These BACs represent collinear regions containing the mapped wheat ESTs and are useful in identifying additional markers for specific wheat chromosome regions (Huo et al. 2006). After Illumina sequencing of wheat chromosome 7BS, assembly was performed by the construction of a syntenic map based on gene order in *B. distachyon*. An earlier reported translocation was delimited and the degree of homoeologous gene conservation between different chromosome arms was analyzed (Berkman et al. 2011, 2012).

Next generation sequencing (NGS) technologies have enabled the rapid generation of large amounts of sequence data. Their recent development has facilitated the analysis of plant species with large genomes, including Triticeae (Metzker 2010). Sequencing of plant genomes is based on BAC-to-BAC sequencing and whole genome shotgun sequencing (Venter et al. 1996). The high content of repetitive DNA in the Triticeae genomes complicates whole genome assembly, especially after shotgun sequencing (Dubcovsky and Dvorak 2007; Luo et al. 2010). *B. distachyon* can

be used as a Pooidae reference physical map, taking advantage of its good assembly. For example, after the NGS of flow sorted wheat 4A chromosomes, syntenic regions were identified in other grass genomes, including *B. distachyon*, for “genome zipper” alignment and genetic map construction. A genome zipper is as an integrated database of known gene indices in syntenic genomes. Gene content, structure of the chromosome, and location of several translocation and inversion events were defined by sequence comparison with other grass genomes including *B. distachyon* (Hernandez et al. 2012). Perhaps the best example of the exploitation of the *B. distachyon* genome in uniting sequencing data, physical maps, genetic maps, and expression data is the recent “gene-ome” of barley (IBSC 2012).

24.5.3 Markers for Cereal Gene Mapping

Currently the whole genome sequence of Bd21 is known and additional sequence data from EST collections and the resequencing of several *B. distachyon* accessions is available. Through its sequence, *B. distachyon* contribute to the generation of genomic markers, which are important for several applications including genetic map construction, map-based cloning of trait-related genes, anchoring the genetic map to the physical map, genomic comparisons, and marker-assisted breeding. For example, “conserved ortholog set” (COS) markers were developed from orthologous genes conserved between rice, *B. distachyon*, sorghum, and wheat (Fulton et al. 2002; Paux et al. 2011). Even SSR markers can be adapted; *B. distachyon* SSRs were adapted to a bioenergy crop, *Miscanthus sinensis* (Zhao et al. 2011).

Recently, a Bd3-1 X Bd1-1 population was used to create an AFLP-based linkage map. Anchoring to the genome sequence was performed with SSR and SNP markers. Three QTLs were found to be involved in resistance of *Brachypodium* to false brome rust (*Puccinia brachypodii*; Barbieri et al. 2012). In another study, markers were developed based on synteny with *Brachypodium* and rice to map powdery mildew resistance gene PmAS846 of wild emmer wheat to 5BL and it was observed that marker order is collinear with genomic regions on *Brachypodium* chromosome 4 (Xue et al. 2012a).

B. distachyon has served for mapping genes giving resistance to other cereal diseases as well. Two quantitative trait loci (QTL) for powdery mildew (*Blumeria graminis*) resistance in barley were fine-mapped using barley markers and other markers developed with comparative genomic analysis of QTLs in other grasses including *Brachypodium*. It was also shown that the mapped regions on chromosome 7A had syntenic regions with *B. distachyon* chromosome 1 (Silvar et al. 2012). For fine mapping of the powdery mildew resistance gene *Pm6* gene in wheat, which was earlier mapped to wheat chromosome 2BL, markers based on collinearity with rice and *B. distachyon* were used. The markers were shown to cover a region syntenic to chromosome 5L of *B. distachyon* and flanking the *Pm6* locus, and an associated LRR-receptor-like protein kinase was identified (Qin et al. 2011). Based on the collinearity of wheat with *Brachypodium* and rice, EST-STS markers were developed and used

to map the powdery mildew resistance gene MIIW170 onto chromosome 2BS of wild emmer wheat. Four resistance gene analog sequences were annotated in the orthologous *B. distachyon* genomic region, which now can be used for map-based cloning of MIIW170 (Liu et al. 2012).

The potent eyespot resistance gene *Pch1* from *Aegilops* was previously introgressed to wheat chromosome 7DL, but the hybrid was limited by the linkage drag of yield limiting traits. Conserved orthologous sequence (COS) co-dominant PCR markers were developed using *B. distachyon* and recombinants were screened on heterozygotes in F₂ populations of wheat and *Aegilops* around 7DL (Burt and Nicholson 2011). Using co-linearity of this wheat chromosome with *B. distachyon* chromosome 1 and other grass chromosomes, *Pch1* was localized to an interval containing candidate gene regions on which map-based cloning can be performed, based on the *B. distachyon* sequence (Burt and Nicholson 2011).

Brachypodium has served also in mapping of diverse traits in addition to disease resistance. In a recent study, seven wheat chromosomal regions involved in grain dietary fiber content were identified. Genes that were differentially expressed during grain development and between genotypes with different grain fiber contents were also detected. Comparative studies identified candidate genes for the trait, based on comparison to *B. distachyon* and other grass genomes (Quraishi et al. 2011). Wheat seed dormancy, earlier shown to be related to a yield QTL, was found to be located on chromosome 2B. Markers linked to this QTL were developed and the region fine-mapped, based on the *B. distachyon* and rice genomes. The region was found to collinear with a region on *B. distachyon* Bd1 (Somyong et al. 2011). One of the loci controlling spike density, dense spike (*dsp*), was mapped to the centromere of chromosome 7H in barley. Comparison with collinear regions of other grasses including *Brachypodium* showed that this region contains more than 800 genes (Shahinnia et al. 2012), illustrating the limits of combining genetics with genomics in low-recombination regions.

B. distachyon BESs from BAC libraries have significant matches to wheat ESTs mapped to individual chromosome bin positions. These BACs represent collinear regions containing the mapped wheat ESTs and have been useful in identifying additional markers for specific wheat chromosome regions (Huo et al. 2006). For *Lolium multiflorum*, a genetic map was saturated with markers using Diversity Array Technology (DARt) markers and the DarTFest array. Comparative analysis of these markers with rice and *B. distachyon* was then performed (Bartos et al. 2011).

For *Leymus*, which is an allotetraploid, genetic maps with linkage groups (LG) including several markers were previously used for mapping QTLs. However, recently, a consensus map was developed with new markers and arranged so that linkage groups can be aligned to Triticeae and *Brachypodium* and homoeologous groups shown. Previously, chromosomes of *Leymus* were transferred to wheat and this study was performed on wheat-*Leymus* chromosome introgression lines. Reciprocal translocations between 4 and 5L in both *Leymus* and *Triticum monococcum* were aligned to regions of *Brachypodium* chromosome 1. Glauousness genes on *Leymus* and wheat chromosome 2 were aligned to a region of *Brachypodium* chromosome 5. The *Leymus* chromosome-2 self-incompatibility gene aligns to *Brachypodium* chromosome 5 (Larson et al. 2012).

24.5.4 Map-based Cloning and Genetic Mapping

Positional gene isolation in unsequenced species generally requires either a reference genome sequence or a reference gene content and order based on conservation of synteny with a genomic model. Due to the lack of a complete reference genome sequence and low gene density in many grass genomes, fine mapping and map-based gene isolation often relies on exploiting conserved synteny with model grass species (e.g. rice and *Brachypodium*). For these purposes, ‘homology bridges’ between the model genome and the target region that contains a gene of interest which are sequences of genetically mapped genes, are necessary and increasing the density of these genes around a target locus is important.

By now, in crops, several traits were fine mapped using *B. distachyon* as a model genome (Turner et al. 2005; Griffiths et al. 2006; Spielmeyer et al. 2008) and in some SSR markers from *B. distachyon* were used in the process (Azhaguvel et al. 2009; Vogel et al. 2009; Garvin et al. 2010). One example is the chromosomal pairing locus, *Ph1*, which was mapped to its location in wheat chromosome 5B using markers from an orthologous region in *B. sylvaticum*, even before the *B. distachyon* sequence was produced (Griffiths et al. 2006). At that time, the *Oryza sativa* genome sequence was available, but mapping could not be done using *O. sativa* markers because its sequence in this region was too divergent from wheat. Other genes that were mapped with *B. distachyon* markers include the *Lr34/Yr18* rust resistance gene in wheat and the *Ppd-H1* photoperiod response gene in barley (Turner et al. 2005; Spielmeyer et al. 2008). However, in a comparative study, the wheat Hessian fly resistance gene H26 was shown not to be conserved in *B. distachyon* and the use of *B. distachyon* in map based gene cloning approaches was questioned (Yu et al. 2009).

24.5.5 Comparative Studies Using Bioinformatics

24.5.5.1 Genome Evolution

The *B. distachyon* genome, with its collinearity and well-annotated genes, is an ideal starting point for structure and function analyses. The *B. distachyon* genome is an enormous aid in predictive assembly of short-read sequences. For example, following Illumina short-read sequencing of wheat chromosome 7BS, assembly was performed by the construction of a syntenic build based on gene order in *B. distachyon* (Berkman et al. 2011, 2012). A previously reported translocation was delimited and the degree of homoeologous gene conservation between different chromosome arms was reported. For chloroplast genomes, SSRs and their flanking regions were analyzed by multiple alignments of several grass species including *B. distachyon* (Melotto-Passarin et al. 2011). This was performed to detect DNA sequence variations and organization of cpSSRs in genic and intergenic regions. It was found that cpSSRs are polymorphic, limited to intergenic regions, and stable through grass family. With this data a plastome database was created.

For nuclear genes there are many examples of *in silico* studies. PMM genes were isolated from different Triticeae species including *Brachypodium* and bioinformatic and biochemical analyses were used to study the evolution of functionality of these genes and proteins among Triticeae. This led to the discovery of a duplication event in the gene prior to bread wheat evolution and the presence of more temperature tolerance of these proteins in *Triticum* compared to *Brachypodium*, which can give an understanding of temperature adaptability in bread wheat (Yu et al. 2010). Likewise, two paralogous plant architecture controlling genes, ABCG5/6 and their gene families were identified and their syntenic relations and functions were characterized based on phylogenetic studies and comparative genomics involving *Brachypodium* (Shinozuka et al. 2011).

Looking instead at the genes that do not fit the genome zipper, it is known that collinearity of genes in plant genomes is inversely proportional to their evolutionary distance. In a recent study, non-collinear genes were identified by comparing genomes of *B. distachyon*, sorghum and rice. This study led to the finding that this disruption of collinearity can result from DSB repair to patch gaps formed by transposon movement (Wicker et al. 2010). In another paper by the same group, the synteny of genes in the previously determined bread wheat Triticeae group 1 syntenic region were assessed in comparison to other crop species and to *B. distachyon*. The results achieved showed that even if the syntenic genes were conserved, there were several other nonsyntenic genes that had their homologs elsewhere in other crops and *B. distachyon*. This showed that the total gene number in bread wheat is overestimated due to pseudogenes resulting from movement of transposable elements and DSB repair (Wicker et al. 2011).

24.5.5.2 Transcription

Analysis of transcriptional patterns in various cereals has been enhanced by use of *B. distachyon* database tools. For example, a transcription map of a wheat chromosome (3B) was made. Based on gene positions in BACs and bins, orthologous genes and their level of synteny were determined in *Brachypodium* and *Oryza* (Paux et al. 2011). Putative functions were assigned to the wheat chromosome 3B unigenes with the Gene Ontology (GO) annotations of *O. sativa* and *B. distachyon* orthologs. Co-expressed and co-functional gene islands were identified and their conservation in rice or *B. distachyon* was studied. Similarly, data from barley expression chips were used to create co-expression clusters, which represent networks of different biological functions and can be used to facilitate gene discovery in barley and other crops (Mochida et al. 2011a). The annotation of these clusters was performed with comparison to genes from other organisms including *B. distachyon*.

In a global analysis, 86 % of the estimated 32,000 barley genes were assigned to individual chromosome arms and assembled into a scaffold of putative linear order (Mayer et al. 2011). For this purpose, a genome zipper for the grasses including *B. distachyon*, was used. In wheat, a transcription map was developed for chromosome 3B, which shows a two-fold increase in the number of genes in islands, resulting in

an increase in gene density towards the telomeres on this chromosome (Rustenholz et al. 2011). The acquiring of a common regulatory pattern during evolution was proposed for these islands since they were found to be co-functional and co-expressed. Comparative analysis of the chromosome with rice and *B. distachyon* showed that gene islands predominantly had genes originating from inter-chromosomal gene duplications and that the co-expressed and co-functional genes were predominantly not conserved, suggesting a recent evolutionary origin.

24.5.5.3 Gene Function

To study past and possible future evolutionary patterns of new functions in plant metabolism, cytochrome P450 (CYP) complements of rice and *B. distachyon* were compared. The results show that evolution of new functions in plant metabolism is a very long term process and highlight convergence of essential functions (Nelson and Werck-Reichhart 2011). Another gene family, the oxidosqualene cyclases, includes enzymes involved in the synthesis of metabolites, the triterpenoid skeletons. In a recent study, they were functionally and evolutionarily analyzed across grasses including *B. distachyon*, showing that the increase in the number of OSCs in higher plants is due to tandem duplication followed by diversifying selection (Xue et al. 2012b). Recently, a phylogenetic, molecular, and comparative analysis at the DNA, protein, and genetic/physical map levels was performed for the cytokinin oxidase/dehydrogenase gene family across the Poaceae including *B. distachyon* (Mameaux et al. 2012). This family is important because OsCKX2 was previously shown to be involved in yield increase.

Transcription factors (TFs) are a popular group of genes on which to carry out *in silico* analyses. A database, the GramineaeTFDB, which contains predictions of crops TFs based on *in silico* analyses of available TF collections, can be used for functional and comparative genomics of TFs. It harbors a tool to search for putative *cis*-elements in the promoter regions of TFs and predict the functions of the TFs (Mochida et al. 2011b). Barley NAC TFs were identified *in silico* and compared to NAC proteins from other grasses including *B. distachyon* to reveal their subfamily membership. Experimental analysis has shown that their functions are conserved among grasses, such as secondary cell wall biosynthesis, leaf senescence, root development, seed development, and hormone regulated stress responses (Christiansen et al. 2011).

In silico microRNA and target identification has likewise been carried out through comparative sequence analysis of related species. Using *in silico* methods, *B. distachyon* microRNAs can be identified and accessed data from MIRBASE (Unver and Budak 2009). Some of the *Brachypodium* microRNA targets identified were found to encode transcription factors regulating plant development, morphology and flowering time and others were involved in stress response.

24.6 Analysis in *Brachypodium* of Traits for Cereal Crops

Brachypodium serves as a useful functional genomics model of cereal crops to quickly determine gene function for a range of important biological traits. In addition to its appropriate biological properties, a great amount of genomic knowledge is present and efficient protocols for *Brachypodium* have been developed, as described above. Functions can be assigned to crop genes with the aid of *Brachypodium* using several approaches. These can range from methods involving bioinformatics based on alignment of crop and *Brachypodium* genomes to approaches involving forward and reverse genetics. In order to assign function to each *Brachypodium* gene, a major aim is to achieve sufficient genome coverage so that each *Brachypodium* gene will have a corresponding mutant among the generated mutant populations. However there are also alternative strategies immediately applicable based on gene silencing.

24.6.1 Biotic Stress Resistance

Nuclear Factor Y (NF-Y) transcription factors are known to be involved in important traits such as drought tolerance, flowering time, and seed development in *Arabidopsis*. The *B. distachyon* NF-Y proteins were identified, annotated, and characterized through phylogenetic and orthology based studies and tissue specific expression patterns (Cao et al. 2011). The *B. distachyon* NF-YP, identified via orthology to *Arabidopsis* floral-promoting NF-Y proteins, was cloned; it rescued the late flowering phenotype of the mutant *Arabidopsis* after transformation. Overall, in this study, it was found that NF-Y was functionally conserved between dicots and monocots in several aspects. *Brachypodium*-based information is more readily translatable to monocot plants (Cao et al 2011).

B. distachyon has been used to study responses of Triticeae to pathogens (Draper et al. 2001). The first related reports were on model interactions of the Triticeae with rust and mildew, on which further research is required (Draper et al. 2001; Ayliffe et al. 2008). Studies on establishing model interactions have been performed for *Puccinia striiformis* (yellow rust), *Puccinia recondita* (brown rust), *Puccinia coronata* (crown rust), *Puccinia brachypodii* (false brome rust), *Puccinia graminis* (stem rust), *Fusarium graminearum*, and *Fusarium culmorum*, head blight (Peraldi et al. 2011; Barbieri et al. 2012). Model interactions have been established also for aphids including *Schizaphis graminum* and *Diuraphis noxia* (Azhaguvel et al. 2009). However, no fungus-host interactions were established for *Blumeria* (powdery mildew), *Septoria* (leaf blotch), *Rhizoctonia solani* (several diseases) or *Gaeumannomyces* sp. (take-all).

Recently, the first quantitative trait locus (QTL) analysis in *B. distachyon* was undertaken for resistance to false brome rust, *Puccinia brachypodii*. This was performed through creating an AFLP-based linkage map on a Bd3-1 X Bd1-1 population, in which the two parental populations were selected based on their differing resistance

to the pathogen. One QTL was shown to affect resistance at both seedling and advanced growth stages and was mapped to chromosome 2, while three gave resistance only to seedlings and resided on chromosomes 3 (2) and 4 (1) respectively (Barbieri et al. 2012).

Additionally, comparative studies of resistance genes have been undertaken. Genetic markers for barley chromosomes were used to analyze the synteny of rust resistance genes; the *Rpg1* and *Rpg4* stem rust genes were found to have orthologs in *B. distachyon*. In another comparative study, the genetic diversity of *B. distachyon* was assessed using EST and microsatellite markers and this information was related to the feeding preferences of the wheat greenbug (*Schizaphis graminum* Rondani) and the Russian wheat aphid (RWA), *Diuraphis noxia* (Azhaguvel et al. 2009). The “enemy release hypothesis”, which suggests that the fitness of a species is greater in an invaded range than in its natural range, was tested on *B. sylvaticum*. The variants and frequency of generalist and specialist pathogens of *Brachypodium*, including insects and fungi, were analyzed in two ranges and in relation to their severity (Halbritter et al. 2012). Nevertheless, *Brachypodium* cannot be used as a model organism in all biotic stress studies for the cereals due to orthology limitations. For example, the Wheat-Induced Resistance 1 (TaWIR1) gene family, which is strikingly induced in response to several pathogens, has homologues in rice, barley, and wheat but not in *B. distachyon* (Tufan et al. 2012).

24.6.2 Abiotic Stress Tolerance

Brachypodium has a good potential to be used as a model in abiotic stress studies because it has a preference for growth at higher altitudes and on marginal ground. There has been some relevant work using *B. pinnatum* and *B. rupestre* (Hurst and John 1999; Liancourt et al. 2005; Matts et al. 2010). Other abiotic stresses that effect crop yields negatively are flooding and drought (Cassman 1999). To create an overall *Brachypodium* expression atlas, including differential expression of genes at various developmental stages, under a variety of stresses, and at different times of the day, a Bd21 genomic sequence based array was developed. In a classification of co-expressed genes in barley from microarray hybridization experiments with the aid of *B. distachyon*, modules involved in drought stress were identified (Mochida et al. 2011a).

In the recent years, microRNAs have been identified in several plants including *Brachypodium* (MIRBASE) and the fluctuations in their expression levels in response to various abiotic stresses studied. Using *in silico* methods, *Brachypodium* microRNAs were identified and can be accessed from MIRBASE. Some of the *Brachypodium* microRNA targets identified were found to be involved in stress response (Unver and Budak 2009). More recently, *B. distachyon* microRNAs responsive to dehydration stress were identified in root and leaf tissues on a microarray platform and validated by qRT-PCR (Budak and Akpinar 2011). Furthermore, targets of these microRNAs were predicted *in silico* and validated with RLM-RACE. Some of the *B. distachyon*

miRNAs were shown to be cold-responsive, though most of these are specific to *Brachypodium*. This suggests *Brachypodium* has some specific mechanisms for cold response.

Nuclear Factor Y (NF-Y) transcription factors are known to be involved in important traits such as drought tolerance, flowering time and seed development in *Arabidopsis*. In a recent study, *B. distachyon* NF-Y proteins were identified, annotated and characterized through phylogenetic and orthology based studies and tissue specific expression patterns. *Brachypodium* NF-YP, identified via orthology to *Arabidopsis* floral-promoting NF-Y proteins, was cloned and rescued the late flowering phenotype of the mutant *Arabidopsis* after transformation. Overall, in this study, it was found that NF-Y was functionally conserved between dicots and monocots in several aspects.

24.6.3 Bioenergy

Herbaceous energy crops among the Poaceae such as *Miscanthus* and switchgrass (*Panicum virgatum*) are potential sources of renewable energy and the subjects of intensive research. However there is a lack of information on the biological basis of bioenergy traits. *Brachypodium* can serve as a good model to study biofuel crops; for this reason, the *B. distachyon* genome sequence was funded in the main by the US Department of Energy. Biofuels from crops are generally based on the production of ethanol or diesel during fermentation; lignocellulosic cell walls can serve as fermentation substrates (Chang 2007; Gomez et al. 2008a, b).

Understanding the construction, degradation, and saccharification of lignocellulose in the cell wall is therefore important for producing biofuel from biomass (Wyman 2007). In a recent study of classification of co-expressed barley genes from chip experiments with the aid of *B. distachyon*, modules involved in cellulose biogenesis were identified (Mochida et al. 2011a). The cell wall composition of *Brachypodium* was shown to be more similar to bioenergy crops and cereals, in comparison to that of *Arabidopsis* (Gomez et al. 2008b; Opanowicz et al. 2008). Additionally, hemicelluloses in the cell walls of *Brachypodium*, *H. vulgare* and *T. aestivum* were compared and similarities and differences were noted (Christensen et al. 2010). A study on the mechanism of cell wall saccharification has been performed on *Brachypodium* stems (Gomez et al. 2008b). Furthermore, it was shown that extraction of sugars can be improved by the modulation of cell wall biosynthesis genes (Van Hulle et al. 2010). *Brachypodium* has also been used with switchgrass in a study of senescence, the delay of which is involved in a sugar level increase and nutrient mobilization before winter (Yang and Ohlrogge 2009).

Analyses on saccharification have also gone forward in *Brachypodium*. Mild acidic conditions was used to analyze the saccharification process during stem hydrolysis, showing the predominance of hydrolysis of hemicellulose hydrolysis in comparison to cellulose, with scanning electron microscopy used to demonstrate the

tendency of fibrils for hydrolysis (Gomez et al. 2008a). New assays are being developed to assess the liability of the plant to fuel formation. Recently such an assay was developed to determine the ethanol production efficiency of *Clostridium phytofermentans* on different plants as related to xylan metabolism (Lee et al. 2012). Natural genetic variation effecting conversion efficiency was characterized in *B. distachyon*.

24.6.4 Yield, Grain Characteristics and Plant Development

Although *Brachypodium* is an undomesticated grass species, it is an appropriate model organism for yield studies because, relative to its stature, it has long grains within a large spike and seeds similar to wheat (Draper et al. 2001; Garvin et al. 2008; Opanowicz et al. 2008). It is thought that some of the first grains to be processed were from *Brachypodium* species (Revedin et al. 2010). *Brachypodium* was shown to contain proteins with similarity to the seed storage protein glutenin (Laudencia-Chinguanco and Vensel 2008; Gu et al. 2010; Larre et al. 2010). *Brachypodium* seed storage proteins were found to be mostly globulins and prolamins. Subcellular localization studies revealed with microscopy showed glutelin bodies within the endosperm (Larre et al. 2010). In another study, grain development and filling was extensively studied in the Bd211 line to show the phases of morphogenesis and patterns of protein, lipid, sugar, and starch accumulation (Guillon et al. 2012). Distinct beta-glucans were found to present in *Brachypodium* compared to other cereals.

Seed and grain dormancy is an important agronomic trait. In a recent report, the dormancy characteristics of different *Brachypodium* genotypes were studied along with the effect of light quality on germination, gene expression, and abscisic acid level differences among dormant and non-dormant genotypes (Barrero et al. 2012). It was found that dormancy and germination were similar in *Brachypodium* to other cereals, which makes it an appropriate model to study these traits. A region important for dormancy in wheat and located on chromosome 2B, which is related to yield, is collinear with a region on *Brachypodium* Bd1 (Somyong et al. 2011).

Traits affecting development and plant architecture are also related to agronomic performance and yield. The Earliness Per Se gene *Eps-1* of *Triticum monococcum* is involved in flowering time and spike development. Its orthologous genes were identified in a recent study in *B. distachyon* and other grasses (Faricelli et al. 2010). Two genes in *B. distachyon* and wheat, *Mot1* and *FtsH4*, were linked to the phenotype and the *Eps-1* position on the T. monococcum genetic maps. *Brachypodium* has also been used as a model in studies examining flowering. *Brachypodium* miRNAs were predicted to target plant transcription factors that regulate development and flowering time (Unver and Budak 2009). To demonstrate the delay of heading, rye *Terminal Flowering 1* genes were expressed in *B. distachyon* (Olsen et al. 2006).

Root systems affect yield by modulating water uptake during flowering and development. *Brachypodium*, with its wheat-like roots and small stature, is a good system in which to phenotype roots and to identify related genes (Chochois et al. 2012). In the root epidermis of Poaceae members, hair cells are smaller than other cells. It was shown that this phenomenon is due to asymmetric cytokinesis in *Brachypodium*, but

not in rice, shedding light on the evolution of this mechanism in the Pooideae (Kim and Dolan 2011).

24.7 Direct Agricultural Use of *Brachypodium*

Brachypodium is very well adapted to the Mediterranean area, where olive is a major crop. Soil erosion is a major environmental problem for the Mediterranean olive groves, and the most efficient and sustainable solution is the use of cover crops among the trees (Pastor et al. 1997; Gómez et al. 2011). Grass covers avoid soil loss while maintaining moisture and nutrients, they contribute to increase the water use efficiency and they facilitate machinery traffic. Recently, *Brachypodium* has been evaluated and selected for grass cover use and, as a result, two commercial varieties have been registered in the EU: Ibros (*B. hybridum*) and Zulema (*B. distachyon*). These were obtained via domestication of natural populations (Soler et al. 2004) and they are mainly grown in Southern Spain olive groves. *Brachypodium* has also been shown a promising soil cover for hillside and steep vineyards (Marques et al. 2010; Ruiz-Colmenero et al. 2011). Among the main advantages of *Brachypodium* as a soil cover in relation to other species are its medium to low stature and its excellent soil cover even in summer because of its persistent stubble, thus facilitating machinery traffic.

24.8 Conclusions

Brachypodium, in particular the diploid annual *B. distachyon*, has matured both as a useful platform for research on the more complex cereal crops and their genomes and as an experimental system in its own right. As a basis for research in the Triticeae and other cereals, its small, well-characterized genome and phylogenetic proximity to the “core Poooids”, particularly the temperate staples barley, wheat, rye, and oat as well as temperate fodder grasses, has proven to be critical. For direct analyses of traits important in the crop plants as well as to examine questions in plant evolution and population genetics, its rapid life cycle, small stature, and general biochemical, physiological, and morphological similarity to cereal crops is very convenient. The germplasm and mutant collections, as well as the closely related *Brachypodium* ploidy series and perennial *B. sylvaticum*, offer paths to expand studies of *B. distachyon* outward in new directions. Although the large and complex genomes of the crop plants themselves are now tractable (IBSC 2012), all of these features of *Brachypodium* as a research system suggest that it will not soon be outmoded.

References

Alves SC, Worland B, Thole V et al (2009) A protocol for *Agrobacterium*-mediated transformation of *Brachypodium distachyon* community standard line Bd21. Nat Protoc 4:638–649

- Ayliffe M, Singh R, Lagudah E (2008) Durable resistance to wheat stem rust needed. *Curr Opin Plant Biol* 11:187–192
- Azhaguvel P, Li W, Rudd JC et al (2009) Aphid feeding response and microsatellite-based genetic diversity among diploid *Brachypodium distachyon* (L.) Beauv accessions. *Plant Genet Resour Charact Util* 7:72–79
- Barbieri M, Marcel TC, Niks RE et al (2012) QTLs for resistance to the false brome rust *Puccinia brachypodii* in the model grass *Brachypodium distachyon* L. *Genome* 55:152–163
- Barrero JM, Jacobsen JV, Talbot MJ et al (2012) Grain dormancy and light quality effects on germination in the model grass *Brachypodium distachyon*. *New Phytol* 193:376–386
- Bartos J, Sandve SR, Kolliker R et al (2011) Genetic mapping of DArT markers in the *Festuca-Lolium* complex and their use in freezing tolerance association analysis. *Theor Appl Genet* 122:1133–1147
- Berkman PJ, Skarshewski A, Lorenc MT et al (2011) Sequencing and assembly of low copy and genic regions of isolated *Triticum aestivum* chromosome arm 7DS. *Plant Biotechnol J* 9:768–775
- Berkman PJ, Skarshewski A, Manoli S et al (2012) Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl Genet* 124:423–432
- Bolot S, Abrouk M, Masood-Quraishi U et al (2009) The ‘inner circle’ of the cereal genomes. *Curr Opin Plant Biol* 12:119–125
- Botstein D, White RL, Skolnick M, Davis RW (1980) Construction of a genetic linkage map using restriction fragment length polymorphisms. *Am J Human Genet* 32:314–331
- Brown TA, Jones MK, Powell W, Allaby RG (2009) The complex origins of domesticated crops in the Fertile Crescent. *Trends Ecol Evol* 24:103–109
- Buchmann JP, Matsumoto T, Stein N et al (2012) Interspecies sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity. *Plant J* 71:550–563
- Budak H, Akpinar A (2011) Dehydration stress-responsive miRNA in *Brachypodium distachyon*: evident by genome-wide screening of microRNAs expression. *OMICS* 15:791–799
- Burt C, Nicholson P (2011) Exploiting co-linearity among grass species to map the *Aegilops ventricosa*-derived *Pch1* eyespot resistance in wheat and establish its relationship to *Pch2*. *Theor Appl Genet* 123:1387–1400
- Cabrera A, Kozik A, Howad W et al (2009) Development and bin mapping of a Rosaceae Conserved Ortholog Set (COS) of markers. *BMC Genomics* 10:562
- Cao S, Kumimoto RW, Siriwardana CL et al (2011) Identification and characterization of NF-Y transcription factor families in the monocot model plant *Brachypodium distachyon*. *PLoS One* 6:e21805
- Cassman KG (1999) Ecological intensification of cereal production systems: yield potential, soil quality, and precision agriculture. *Proc Natl Acad Sci U S A* 96:5952–5959
- Catalan P (2003) *Brachypodium*. In *Catalogue of New World Grasses (Poaceae): IV. Subfamily Pooideae*. *Contr U.S. Natl Herb* 48:143–145
- Catalán P, Olmstead RG (2000) Phylogenetic reconstruction of the genus *Brachypodium* P-Beauv. (Poaceae) from combined sequences of chloroplast *ndhF* gene and nuclear ITS. *Plant Syst Evol* 220:1–19
- Catalán P, Kellogg EA, Olmstead RG (1997) Phylogeny of Poaceae subfamily Pooideae based on chloroplast *ndhF* gene sequences. *Mol Phylogenet E* vol 8:150–166
- Catalán P, Muller J, Hasterok R et al (2012) Evolution and taxonomic split of the model grass *Brachypodium distachyon*. *Ann Bot* 109:385–405
- Chang MC (2007) Harnessing energy from plant biomass. *Curr Opin Chem Biol* 11:677–684
- Chochois V, Vogel JP, Watt M (2012) Application of *Brachypodium* to the genetic improvement of wheat roots. *J Exp Bot* 63:3467–3474
- Christiansen P, Andersen CH, Didion T et al (2005) A rapid and efficient transformation protocol for the grass *Brachypodium distachyon*. *Plant Cell Rep* 23:751–758

- Christensen U, Alonso-Simon A, Scheller HV et al (2010) Characterization of the primary cell walls of seedlings of *Brachypodium distachyon*—a potential model plant for temperate grasses. *Phytochemistry* 71:62–69
- Christiansen MW, Holm PB, Gregersen PL (2011) Characterization of barley (*Hordeum vulgare* L.) NAC transcription factors suggests conserved functions compared to both monocots and dicots. *BMC Res Notes* 4:302
- Dalmais M, Schmidt J, Le SC (2008) UTILdb, a *Pisum sativum* in silico forward and reverse genetics tool. *Genome Biol* 9:R43
- Demircan T, Akkaya MS (2010) Virus induced gene silencing in *Brachypodium distachyon*, a model organism for cereals. *Plant Cell Tissue Organ Culture* 100:91–96
- Distelfeld A, Ezrati S, Eilam T et al (2011) Characterization of the diploid and tetraploid *Brachypodium distachyon* populations in Israel. First European *Brachypodium* Workshop. October 19–21 2011. Versailles Cedex, INRA, France. List of abstracts: https://colloque4.inra.fr/1st_european_Brachypodium_workshop/List-of-abstracts
- Drader T, Kleinhofs A (2010) A synteny map and disease resistance gene comparison between barley and the model monocot *Brachypodium distachyon*. *Genome* 53:406–417
- Draper J, Mur LAJ, Jenkins G et al (2001) *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant Physiol* 127:1539–1555
- Dubcovsky J, Dvorak J (2007) Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science* 316:1862–1866
- Ellwood SR, Phan HT, Jordan M et al (2008) Construction of a comparative genetic map in faba bean (*Vicia faba* L.); conservation of genome structure with *Lens culinaris*. *BMC Genomics* 9:380
- Faricelli ME, Valarik M, Dubcovsky J (2010) Control of flowering time and spike development in cereals: the earliness *per se* *Eps-1* region in wheat, rice, and *Brachypodium*. *Funct Integr Genomics* 10:293–306
- Farrar K, Donnison IS (2007) Construction and screening of BAC libraries made from *Brachypodium* genomic DNA. *Nat Protoc* 2:1661–1674
- Filiz E, Ozdemir BS, Budak H et al (2009) Molecular, morphological, and cytological analysis of diverse *Brachypodium distachyon* inbred lines. *Genome* 52:876–890
- Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14:1457–1467
- Garvin DF, Gu YQ, Hasterok R et al (2008) Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. *Crop Sci* 48:S69–84
- Garvin DF, McKenzie N, Vogel JP et al (2010) An SSR-based genetic linkage map of the model grass *Brachypodium distachyon*. *Genome* 53:1–13
- Giraldo P, Rodríguez-Quijano M, Vázquez JF et al (2012) Validation of microsatellite markers for cytotype discrimination in the model grass *Brachypodium distachyon*. *Genome* 55:523–527
- Gómez JA, Llewellyn C, Basch et al (2011) The effects of cover crops and conventional tillage on soil and runoff loss in vineyards and olive groves in several Mediterranean countries. *Soil Use Manag* 27:502–514
- Gomez LD, Bristow JK, Statham ER, McQueen-Mason SJ (2008a) Analysis of saccharification in *Brachypodium distachyon* stems under mild conditions of hydrolysis. *Biotechnol Biofuels* 1:15
- Gomez LD, Steele-King CG, McQueen-Mason SJ (2008b) Sustainable liquid biofuels from biomass: the writing's on the walls. *New Phytol* 178:473–485
- Griffiths SR, Sharp R, Foote TN et al (2006) Molecular characterization of *Ph1* as a major chromosome pairing locus in polyploid wheat. *Nature* 439:749–752
- Gu YQ, Ma Y, Huo N et al (2009) A BAC-based physical map of *Brachypodium distachyon* and its comparative analysis with rice and wheat. *BMC Genomics* 10:496
- Gu YQ, Wanjugi H, Coleman-Derr D et al (2010) Conserved globulin gene across eight grass genomes identify fundamental units of the loci encoding seed storage proteins. *Funct Integr Genomics* 10:111–122

- Guillon F, Larre C, Petipas F et al (2012) A comprehensive overview of grain development in *Brachypodium distachyon* variety Bd21. *J Exp Bot* 63:739–755
- Halbritter AH, Carroll GC, Gusewell S, Roy BA (2012) Testing assumptions of the enemy release hypothesis: generalist versus specialist enemies of the grass *Brachypodium sylvaticum*. *Mycologia* 104:34–44
- Hammami R, Jouve N, Cuadrado A et al (2011) Prolamin storage proteins and allopolyploidy in wild populations of the small grass *Brachypodium distachyon* (L.) P Beauv. *Plant Systematics Evolut* 297:99–111
- Hasterok R, Draper J, Jenkins G (2004) Laying the cytotaxonomic foundations of a new model grass, *Brachypodium distachyon* (L.) Beauv. *Chromosome Res* 12:397–403
- Hernandez P, Martis M, Dorado G et al (2012) Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *The Plant J* 69:377–386
- Holzberg S, Brosio P, Gross C, Pogue GP (2002) Barley stripe mosaic virus-induced gene silencing in a monocot plant. *The Plant J* 30:315–327
- Huo N, Gu YQ, Lazo GR et al (2006) Construction and characterization of two BAC libraries from *Brachypodium distachyon*, a new model for grass genomics. *Genome* 49:1099–1108
- Huo N, Lazo GR, Vogel JP et al (2008) The nuclear genome of *Brachypodium distachyon*: analysis of BAC end sequences. *Funct Integr Genomics* 8:135–147
- Huo N, Vogel JP, Lazo GR et al (2009) Structural characterization of *Brachypodium* genome and its syntenic relationship with rice and wheat. *Plant Mol Biol* 70:47–61
- Huo N, Garvin DF, You FM et al (2011) Comparison of a high-density genetic linkage map to genome features in the model grass *Brachypodium distachyon*. *Theor Appl Genet* 123:455–464
- Hurst A, John E (1999) The biotic and abiotic changes associated with *Brachypodium pinnatum* dominance in chalk grassland in south-east England. *Biol Conservation* 88:75–84
- Jakob SS, Ihlow A, Blattner FR (2007) Combined ecological niche modelling and molecular phylogeography revealed the evolutionary history of *Hordeum marinum* (Poaceae)—niche differentiation, loss of genetic diversity, and speciation in Mediterranean Quaternary refugia. *Mol Ecol* 16:1713–1727
- International Barley Genome Sequencing Consortium (IBSC) (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491:711–716. doi:10.1038/nature11543
- International *Brachypodium* Initiative (IBI) (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* 463:763–768
- Kim CM, Dolan L (2011) Root hair development involves asymmetric cell division in *Brachypodium distachyon* and symmetric division in *Oryza sativa*. *New Phytol* 192:601–610
- Langridge P, Paltridge N, Fincher G (2006) Functional genomics of abiotic stress tolerance in cereals. *Brief Funct Genomic Proteomic* 4:343–354
- Larre C, Penninck S, Bouchet B et al (2010) *Brachypodium distachyon* grain: identification and subcellular localization of storage proteins. *J Exp Bot* 61:1771–1783
- Larson SR, Kishii M, Tsujimoto H et al (2012) *Leymus* EST linkage maps identify 4NsL-5NsL reciprocal translocation, wheat-*Leymus* chromosome introgressions, and functionally important gene loci. *Theor Appl Genet* 124:189–206
- Laudencia-Chinguanco DL, Vensel WH (2008) Globulins are the main seed storage proteins in *Brachypodium distachyon*. *Theor Appl Genet* 117:555–563
- Lee SJ, Warnick TA, Pattathil S et al (2012) Biological conversion assay using *Clostridium phytofermentans* to estimate plant feedstock quality. *Biotechnol Biofuels* 5:5
- Liancourt P, Corcket E, Michalet R (2005) Stress tolerance abilities and competitive responses in a watering and fertilization field experiment. *J Vegetation Sci* 16:713–722
- Liu Z, Zhu J, Cui Y et al (2012) Identification and comparative mapping of a powdery mildew resistance gene derived from wild emmer (*Triticum turgidum* var. *dicoccoides*) on chromosome 2BS. *Theor Appl Genet* 124:1041–1049
- Luo MC, Ma Y, You FM et al (2010) Feasibility of physical map construction from fingerprinted bacterial artificial chromosome libraries of polyploid plant species. *BMC Genomics* 11:122

- Mameaux S, Cockram J, Thiel T et al (2012) Molecular, phylogenetic and comparative genomic analysis of the cytokinin oxidase/dehydrogenase gene family in the Poaceae. *Plant Biotechnol J* 10:67–82
- Manzaneda AJ, Rey PJ, Bastida JM et al (2012) Environmental aridity is associated with cytotype segregation and polyploidy occurrence in *Brachypodium distachyon* (Poaceae). *New Phytol* 193:797–805
- Marques MJ, García-Muñoz S, Muñoz-Organero G, Bienes R (2010) Soil conservation beneath grass cover in hillside vineyards under Mediterranean climatic conditions (Madrid, Spain). *Land Degradation Development* 21:122–131
- Matts J, Jagadeeswaran G, Roe BA, Sunkar R (2010) Identification of microRNAs and their targets in switchgrass, a model biofuel plant species. *J Plant Phys* 167:896–904
- Mayer KF, Martis M, Hedley PE et al (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 23:1249–1263
- McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeted screening for induced mutations. *Nat Biotechnol* 18:455–457
- Melotto-Passarin DM, Tambarussi EV, Dressano K et al (2011) Characterization of chloroplast DNA microsatellites from *Saccharum* spp and related species. *Genet Mol Res* 10:2024–2033
- Metzker ML (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46
- Meyers BC, Kozik A, Griego A et al (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *Plant Cell* 15:809–834
- Mochida K, Uehara-Yamaguchi Y, Yoshida T et al (2011a) Global landscape of a co-expressed gene network in barley and its application to gene discovery in Triticeae crops. *Plant Cell Physiol* 52:785–803
- Mochida K, Yoshida T, Sakurai T et al (2011b) In silico analysis of transcription factor repertoires and prediction of stress-responsive transcription factors from six major gramineae plants. *DNA Res* 18:321–332
- Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution. Grasses, line up and form a circle. *Curr Biol* 5:737–739
- Mur LA, Allainguillaume J, Catalán P et al (2011) Exploiting the *Brachypodium* Tool Box in cereal and grass research. *New Phytol* 191:334–347
- Neji M, Rahmouni S, Saoudi W et al (2011) Morpho-phenologic diversity among Tunisian populations of *Brachypodium distachyon*. First European *Brachypodium* Workshop. October 19-21 2011. Versailles Cedex, INRA, France. List of abstracts: https://colloque4.inra.fr/1st_european_Brachypodium_workshop/List-of-abstracts
- Nelson D, Werck-Reichhart D (2011) A P450-centric view of plant evolution. *Plant J* 66:194–211
- Olsen P, Lenk I, Jensen CS et al (2006) Analysis of two heterologous flowering genes in *Brachypodium distachyon* demonstrates its potential as a grass model plant. *Plant Sci* 170:1020–1025
- Opanowicz M, Vain P, Draper J et al (2008) *Brachypodium distachyon*: making hay with a wild grass. *Trends Plant Sci* 13:172–177
- Ozdemir BS, Hernandez P, Filiz E, Budak H (2008) *Brachypodium* genomics. *Int J Plant Genomics* 2008:536104
- Pacak A, Geisler K, Jorgensen B et al (2010) Investigations of barley stripe mosaic virus as a gene silencing vector in barley roots and in *Brachypodium distachyon* and oat. *Plant Methods* 6:26–33
- Pacurar DI, Thordal-Christensen H, Nielsen KK, Lenk I (2008) A high-throughput Agrobacterium-mediated transformation system for the grass model species *Brachypodium distachyon* L. *Transgenic Res* 17:965–975
- Pastor M, Castro J, Humanes MD, Saavedra M (1997) La erosión y el olivar: Cultivo con cubierta vegetal. Comunicación I+D agroalimentaria. Junta de Andalucía. Consejería de Agricultura y Pesca
- Paux E, Sourdille P, Mackay I, Feuillet C (2011) Sequence-based marker development in wheat: Advances and applications to breeding. *Biotechnol Adv* 30:1071–1088

- Peraldi A, Beccari G, Steed A, Nicholson P (2011) *Brachypodium distachyon*: a new pathosystem to study *Fusarium* head blight and other *Fusarium* diseases of wheat. *BMC Plant Biol* 11:100
- Pérez-Jiménez M, Budak H, Alcaide B et al (2009) Developing a multiplexed set of SSR markers for the analysis of genetic resources in *Brachypodium*. ITMI (International Triticeae Mapping Initiative)/COST Action Tritigen Joint Workshop, Clermont-Ferrand, France, Book of abstracts p. 124
- Pont C, Murat F, Confolent C et al (2011) RNA-seq in grain unveils fate of neo- and paleopolyploidization events in bread wheat (*Triticum aestivum* L.). *Genome Biol* 12:R119
- Qin B, Cao A, Wang H et al (2011) Collinearity-based marker mining for the fine mapping of Pm6, a powdery mildew resistance gene in wheat. *Theor Appl Genet* 123:207–218
- Quraishi UM, Murat F, Abrouk M et al (2011) Combined meta-genomics analyses unravel candidate genes for the grain dietary fiber content in bread wheat (*Triticum aestivum* L.). *Funct Integr Genomics* 11:71–83
- Revedin A, Aranguren B, Becattini R et al (2010) Thirty thousand-year-old evidence of plant food processing. *Proc Natl Acad Sci U S A* 107:18815–18819
- Robertson IH (1981) Chromosome numbers in *Brachypodium* Beauv. (Gramineae). *Genetica* 56:55–60
- Rosegrant MW, Cline SA (2003) Global food security: challenges and policies. *Science* 302:1917–1919
- Routledge AP, Shelley G, Smith JV et al (2004) *Magnaporthe grisea* interactions with the model grass *Brachypodium distachyon* closely resemble those with rice (*Oryza sativa*). *Mol Plant Pathol* 5:253–265
- Rudd S, Schoof H, Mayer KF (2005) PlantMarkers—a database of predicted molecular markers from plants. *Nucleic Acids Res* 33:D628–632
- Ruiz-Colmenero M, Bienes R, Marques MJ (2011) Soil and water conservation dilemmas associated with the use of green cover in steep vineyards. *Soil and Tillage Research* 117:211–223
- Rustenholtz C, Choulet F, Laugier C et al (2011) A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiol* 157:1596–1608
- Salse J, Abrouk M, Bolot S et al (2009) Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci USA* 106:14908–14913
- Schippmann U (1991) Revision der europäischen Arten der Gattung *Brachypodium* Palisot de Beauvois (Poaceae). *Boissiera* 45:1–250
- Schwartz CJ, Doyle MR, Manzaneda AJ et al (2010) Natural variation of flowering time and vernalization responsiveness in *Brachypodium distachyon*. *Bioenergy Res* 3:38–46
- Scofield SR, Huang L, Brandt AS, Gill BS (2005) Development of a virus-induced gene-silencing system for hexaploid wheat and its use in functional analysis of the Lr21-mediated leaf rust resistance pathway. *Plant Physiol* 138:2165–2173
- Shahinnia F, Druk A, Franckowiak J et al (2012) High resolution mapping of Dense spike-ar (dsp.ar) to the genetic centromere of barley chromosome 7H. *Theor Appl Genet* 124:373–384
- Shinozuka H, Cogan NO, Spangenberg GC, Forster JW (2011) Comparative Genomics in Perennial Ryegrass (*Lolium perenne* L.): Identification and Characterisation of an Orthologue for the Rice Plant Architecture-Controlling Gene OsABCG5. *Int J Plant Genomics* 2011:291563
- Silvar C, Perovic D, Scholz U et al (2012) Fine mapping and comparative genomics integration of two quantitative trait loci controlling resistance to powdery mildew in a Spanish barley landrace. *Theor Appl Genet* 124:49–62
- Soler C, Casanova C, Rojo A (2004) Desarrollo de cubiertas vegetales a partir de gramíneas seleccionadas, para su explotación en tierras de olivar. *Actas Hortic* 41:97–100
- Somyong S, Munkvold JD, Tanaka J et al (2011) Comparative genetic analysis of a wheat seed dormancy QTL with rice and *Brachypodium* identifies candidate genes for ABA perception and calcium signaling. *Funct Integr Genomics* 11:479–490

- Sonah H, Deshmukh RK, Sharma A et al (2011) Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. PLoS One 6:e21298
- Spielmeier W, Singh RP, McFadden H et al (2008) Fine scale genetic and physical mapping using interstitial deletion mutants of *Lr34/Yr18*: a disease resistance locus effective against multiple pathogens in wheat. Theor Appl Genet 116:481–490
- Stace CA (2010) New flora of the British Isles, 3rd edn. Cambridge University Press, Cambridge
- Taheri A, Robinson SJ, Parkin I, Gruber MY (2012) Revised selection criteria for candidate restriction enzymes in genome walking. PLoS One 7:e35117
- Talame V, Bovina R, Sanguineti MC et al (2008) TILLMore, a resource for the discovery of chemically induced mutants in barley. Plant Biotechnol J 6:477–485
- Tan G, Gao Y, Shi M et al (2005) SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. Nucleic Acids Res 33:e122–e122
- Thole V, Vain P (2012) Agrobacterium-mediated transformation of *Brachypodium distachyon*. Methods Mol Biol 847:137–149
- Thole V, Alves SC, B W (2009) A protocol for efficiently retrieving and characterizing flanking sequence tags (FSTs) in *Brachypodium distachyon* T-DNA insertional mutants. Nat Protoc 4:650–661
- Thole V, Worland B, Wright J et al (2010) Distribution and characterization of more than 1000 T-DNA tags in the genome of *Brachypodium distachyon* community standard line Bd21. Plant Biotechnol J 8:734–747
- Thole V, Peraldi A, Worland B et al (2012) T-DNA mutagenesis in *Brachypodium distachyon*. J Exp Bot 63:567–576
- Tufan HA, McGrann GR, Maccormack R, Boyd LA (2012) TaWIR1 contributes to post-penetration resistance to *Magnaporthe oryzae*, but not *Blumeria graminis* f. sp. *tritici*, in wheat. Mol Plant Pathol 13:653–665
- Tuna M, Nizam İ, Öney S et al (2011). Genetic Characterization of New *Brachypodium distachyon* Populations from Diverse Geographic Regions in Turkey. First European Brachypodium Workshop. October 19-21 2011. Versailles Cedex, INRA, France. List of abstracts: https://colloque4.inra.fr/1st_european_Brachypodium_workshop/List-of-abstracts
- Turner A, Beales J, Faure S et al (2005) The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. Science 310:1031–1034
- Unver T, Budak H (2009) Conserved microRNAs and their targets in model grass species *Brachypodium distachyon*. Planta 230:659–669
- Vain P, Thole V, Worland B et al (2011) A T-DNA mutation in the RNA helicase eIF4A confers a dose-dependent dwarfing phenotype in *Brachypodium distachyon*. Plant J 66:929–940
- Vain P, Worland B, Thole V et al (2008) Agrobacterium-mediated transformation of the temperate grass *Brachypodium distachyon* (genotype Bd21) for T-DNA insertional mutagenesis. Plant Biotechnol J 6:236–245
- Van Hulle S, Roldan-Ruiz I, Van Bockstaele E, Muylle H (2010) Functional analysis of genes involved in cell wall biosynthesis of the model species *Brachypodium distachyon* to improve saccharification. Sustain Use Genet Divers Forage Turf Breed 5:479–482
- Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364–366
- Vogel J, Hill T (2008) High-efficiency Agrobacterium-mediated transformation of *Brachypodium distachyon* inbred line Bd21-3. Plant Cell Rep 27:471–478
- Vogel JP, Gu YQ, Twigg P et al (2006) EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*. Theor Appl Genet 113:186–195
- Vogel JP, Tuna M, Budak H et al (2009) Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. BMC Plant Biol 9:88
- Wang H, Fang J, Liang C et al (2011) Computation-assisted SiteFinding-PCR for isolating flanking sequence tags in rice. Biotechniques 51:421–423

- Wicker T, Keller B (2007) Genome-wide comparative analysis of *cop* retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual *cop* families. *Genome Res* 17:1072–1081
- Wicker T, Buchmann JP, Keller B (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res* 20:1229–1237
- Wicker T, Mayer KF, Gundlach H et al (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* 23:1706–1718
- Wolny E, Hasterok R (2009) Comparative cytogenetic analysis of the genomes of the model grass *Brachypodium distachyon* and its close relatives. *Ann Bot* 104:873–881
- Wu F, Eannetta NT, Xu Y et al (2009) A COSII genetic map of the pepper genome provides a detailed picture of synteny with tomato and new insights into recent chromosome evolution in the genus *Capsicum*. *Theoretical Appl Genet* 118:1279–1293
- Wyman CE (2007) What is (and is not) vital to advancing cellulosic ethanol. *Trends Biotechnol* 25:153–157
- Xu G, Ma H, Nei M, Kong H (2009) Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proc Natl Acad Sci U S A* 106:835–840
- Xue F, Ji W, Wang C et al (2012a) High-density mapping and marker development for the powdery mildew resistance gene PmAS846 derived from wild emmer wheat (*Triticum turgidum* var. *dicoccoides*). *Theor Appl Genet* 124:1549–1560
- Xue Z, Duan L, Liu D et al (2012b) Divergent evolution of oxidosqualene cyclases in plants. *New Phytol* 193:1022–1038
- Yang Z, Ohlrogge JB (2009) Turnover of fatty acids during natural senescence of Arabidopsis, *Brachypodium*, and Switchgrass and in Arabidopsis beta-oxidation mutants. *Plant Physiol* 150:1981–1989
- Youens-Clark K, Buckler E, Casstevens T et al (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res* 39:D1085–1094
- Yu GT, Cai X, Harris MO et al (2009) Saturation and comparative mapping of the genomic region harboring Hessian fly resistance gene H26 in wheat. *Theor Appl Genet* 118:1589–1599
- Yu C, Li Y, Li B et al (2010) Molecular analysis of phosphomannomutase (PMM) genes reveals a unique PMM duplication event in diverse Triticeae species and the main PMM isozymes in bread wheat tissues. *BMC Plant Biol* 10:214
- Yuan C, Li C, Yan L et al (2011) A high throughput barley stripe mosaic virus vector for virus induced gene silencing in monocots and dicots. *PLoS One* 6:e26468
- Zhao H, Yu J, You FM et al (2011) Transferability of microsatellite markers from *Brachypodium distachyon* to *Miscanthus sinensis*, a potential biomass crop. *J Integr Plant Biol* 53:232–245

Chapter 25

Mining Natural Variation for Maize Improvement: Selection on Phenotypes and Genes

Shilpa Sood, Sherry Flint-Garcia, Martha C. Willcox and James B. Holland

Contents

25.1	Maize History and Classification	617
25.2	Breeding to Enhance Genetic Diversity in Elite Materials	620
25.3	QTL Analysis and its Discontents	621
25.4	Association Analysis	623
25.5	Linkage and Association Analysis in Nested Association Mapping Populations	625
25.6	QTL Fine-Mapping	628
25.7	Marker-Based Selection for Complex Traits in Maize	630
25.8	GEM Allelic Diversity Project	634
25.9	Seeds of Discovery—Large-scale Genotyping and Phenotyping of CIMMYT Germplasm	634
25.10	Bridging the Domestication Bottleneck with Teosinte Introgression Libraries	637
	References	640

Abstract Maize is highly genetically and phenotypically diverse. Tropical maize and teosinte are important genetic resources that harbor unique alleles not found in temperate maize hybrids. To access these resources, breeders must be able to extract favorable unique alleles from tropical maize and teosinte from their population

J. B. Holland (✉) · S. Sood
Department of Crop Science, North Carolina State University,
Raleigh, NC 27695, USA
e-mail: james_holland@ncsu.edu

J. B. Holland
USDA-ARS Plant Science Research Unit, Raleigh, NC 27695, USA

S. Sood
e-mail: shilpa_ood@ncsu.edu

S. Flint-Garcia
USDA-ARS Plant Genetics Research Unit, Columbia, MO 65211, USA

Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA

M. C. Willcox
Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT),
Texcoco, Edo. de México, México

genomic context, where they are linked with many undesired alleles that confer adaptation to tropical environments, ancient farming methods, or wild growth habit (in the case of teosinte). Long-term traditional breeding efforts have demonstrated the value of diverse germplasm to improve maize productivity, while also enhancing the genetic base of cultivated varieties. Genomics provides new opportunities to identify the genes affecting important agronomic traits and to estimate the wide range of allelic effects at such genes. New approaches to complex trait analysis, including joint multiple population analysis, genome-wide association analysis, and genomic selection, can leverage high throughput sequencing and genotyping technologies to improve our understanding of the genome-wide distribution of allele effects across the wide genetic variation in the primary gene pool of maize. Implementing this information for practical maize improvement remains a challenge.

Keywords Maize · Teosinte · Allelic effect · Genome sequencing · Genome-wide association analysis · Linkage drag · Genomic selection · SNP · Candidate gene · Haplotype · Adaptation · Productivity · *Zea mays*

Maize (*Zea mays* L. subsp. *mays*) is an extremely genetically variable crop, adapted to a wide range of habitats, from latitude 40° S to 58° N and including the tropics (Mangelsdorf 1974). In México alone, maize is adapted to environments from 0 to 2900 masl and with 426–4245 mm annual rainfall (Ruiz et al. 2008). The wide genetic variation and adaptation of maize is reflected in its amazing phenotypic diversity for many morphological, developmental, agronomic, and reproductive traits (Kuleshov 1933).

Maize was presumably originally domesticated 5–10,000 years ago in or near Southern México from a progenitor similar to the extant wild teosinte, *Z. mays* subsp. *parviglumis*, hereafter *parviglumis* (Matsuoka et al. 2002). Stringent selection for rare combinations of mutations in a relatively small number of key domestication loci in the earliest phase of the domestication process, followed by thousands of generations of artificial selection for increased ear size and kernel production per plant subjected maize to a population bottleneck, reducing its genetic diversity relative to teosinte (Doebley 2004; Wright et al. 2005). Nevertheless, modern maize retains higher sequence diversity than humans or *Drosophila* (Tenaillon et al. 2001). The predominantly outcrossing mating system of maize, its exposure to selection for adaptation in very diverse environments and for distinct purposes, and the potential for gene flow between maize and its sympatric wild relatives near its center of origin in México all contributed to the relatively high genetic diversity within maize compared to other crops.

Breeders would like to exploit this substantial genetic variation for the purpose of improving elite maize hybrids for important agronomic traits including grain yield and quality, disease and insect resistance, and abiotic stress resistance. Historically, breeders attempted to measure, classify, and exploit maize genetic variation based on observable phenotypic variation. Maize geneticists pioneered the development of molecular marker systems in plants, which provide a means to directly assay genetic variation. In recent years, the ability to characterize genetic variation in maize at the

sequence level has improved dramatically, permitting unprecedented opportunities to identify specific genes (or non-coding sequences) controlling phenotypic variation, and to expose the underlying allelic to direct genic selection.

This review of the use of natural genetic variation in maize will follow the historical development of methodological approaches, from strictly phenotype-based evaluation, classification, and selection, the successes and failures of which are well documented, to current approaches based on gene identification and allele mining, which are just beginning to be tested. Rather than ignore decades of work on phenotype-based breeding with diverse maize, we believe that lessons learned from this research provide a useful framework for considering the likely advantages and disadvantages of modern gene-based selection.

25.1 Maize History and Classification

Maize spread through the Americas following its domestication in Southern México approximately 5,000–10,000 years ago (Matsuoka et al. 2002; Piperno et al. 2009; Van Heerwaarden et al. 2011), resulting in a distribution ranging from the Gaspé Peninsula in modern day Canada ($> 40^\circ$ N) to Chile and Argentina (nearly 40° S) before the arrival of Columbus (Weatherwax 1954). The spread of maize northward from its center of origin has been studied in some detail, revealing that maize was grown east of the Mississippi River by about 2000 years BP ago (Crawford et al. 2006), but that it remained a minor component of the early agricultural system in this area until about 1200 years BP (Smith 1989). A dramatic shift to a maize-based agriculture in North America occurred between 1,200 and 900 years BP; the evolution of the early maturing Northern Flint type was likely an important component of this transition, but the biological changes of maize occurred within dramatic cultural changes that happened during this time. The subsequent Colombian exchange (Crosby 1972) resulted in the relatively rapid dissemination of maize to Europe, followed by Asia and Africa. The natural selection for adaptation to widely diverse ecological habitats combined with artificial selection for human food and ceremonial uses (Weatherwax 1954; Hernández 1985) was the basis for the fantastic display of phenotypic diversity among maize varieties from around the world (Fig. 25.1).

The tremendous variability within maize was recognized early on, and initial attempts to classify the different types of maize were rather artificial, focusing on endosperm type (Sturtevant 1899). Anderson and Cutler (1942) introduced the concept of maize races as a way to delineate groups in which individuals have “a significant number of genes in common,” and suggested some characteristics of the reproductive organs (tassel and ear) that would be useful for grouping maize into races. Such methods, along with geographic origins, informed the large-scale efforts to collect and classify the landraces grown in Latin America in the 1940s and 1950s (Goodman and Brown 1988). This classification effort was performed more or less independently for each country or region, such that relationships among maize populations from different areas were not formally considered, although some race names



Fig. 25.1 A small sample of the phenotypic variability among Latin American races of maize for ear and kernel morphology. Each ear represents a different race, grown under common conditions in a winter nursery in Homestead, FL by Dr. M.M. Goodman. (Photographs by Dr. Jesús Sánchez-Gonzalez)

were used for maize found in different regions. Furthermore, the classifications were performed on a somewhat *ad hoc* basis, without formally defining what “significant number of genes” or character similarity was sufficient to define a race; the authors intended these racial groupings as only preliminary steps in classification of maize (Holland and Nelson 2010). About 250 races have been named for maize of the Americas (M.M. Goodman, pers. comm.), and collection and classification efforts continue to this day in regions where traditional landraces are still grown in México (Ron Parra et al. 2006; Rincón et al. 2010).

Goodman and colleagues formalized the classification of maize and studied relationships among races from different countries using numerical taxonomy (reviewed in Goodman and Brown 1988; Holland and Nelson 2010). The development of isozymes as a genetic marker system in maize provided geneticists with a method to measure relationships among groups without the confounding influence of environment on phenotypic characters. A series of studies by Goodman and colleagues (Goodman and Stuber 1983; Doebley et al. 1984; Doebley et al. 1985; Bretting et al. 1987; Doebley et al. 1988; Bretting et al. 1990; Sanchez and Goodman 1992a, b; Sanchez et al. 2000a, b, 2006, 2007) measured the genetic variation at neutral isozyme loci among and within landraces from the Americas. A key finding of these studies was the very high level of genetic variation within accessions (generally

representing a sample of ears from a single field or village) and races. Genetic differentiation among accessions within a race or among races tends to be low (typically less than 20%), indicating that races and accession groupings account for only a limited amount of genetic variation; the remaining bulk of genetic variation can be found within collections (Sanchez et al. 2000a, b). Where genetic variation follows racial groupings, it is often strongly associated with geography and ecology, altitude in particular (Bretting et al. 1990; Sanchez et al. 2000a). Further, races differ for the amount of variation they contain, with widespread races, particularly those from Mesoamerica, possessing more alleles per locus than races used as specialty varieties and with restricted geographic ranges (Sanchez et al. 2000a). Finally, rare alleles are exceedingly common: 65% of alleles had frequency of 1% or less in the Mexican races analyzed by Sanchez et al. (2000a, b). Reif et al. (2006) largely confirmed these findings with SSR analyses of Mexican landraces. Pressoir and Berthaud (2004a, b) measured both SSR and trait variation within and among landrace samples collected from a small region of México, finding strong differentiation among populations from different villages for certain ear traits, but almost no differentiation for random SSR markers. They interpreted these apparently contradictory results as evidence that gene flow is very common among villages (facilitated by regular seed exchanges), reducing differentiation for most of the genome, but that strong divergent local selections for specific traits result in differentiation at those loci controlling the targeted traits.

SSR evaluations of maize landraces from throughout the Americas indicate that at the broadest scale, American landraces can be grouped into four geographically-based clusters: highland Mexican, Northern United States, lowland tropical, and highland Andean (Vigouroux et al. 2008). Landraces from some geographic areas represent mixtures of these mega-groupings: e.g., Southeastern USA landraces appear to have originated from a mixture of Northern USA and tropical lowland types, whereas lowland Brazilian maize appears to have arisen from admixture between Andean and tropical lowland groups. Variation among landraces within these mega groups is highest for Mexican and lowest for the Andean and Northern US landraces, which represent the extremes of geographic spread from the center of origin. SSR studies also clarified the relationships between landraces of Europe and the Americas, suggesting two distinct introductions of maize to Europe: first by Caribbean maize and later by Northern Flint types (Rebourg et al. 2003; Dubreuil et al. 2006).

In contrast to the high levels of molecular variation observed in landraces and tropical germplasm in general, modern temperate hybrids exhibit high degrees of relatedness arising from the use of a limited set of founder lines (Smith et al. 1992; Duvick et al. 2004). A 23% reduction in sequence variation was observed between landraces and public USA inbreds (Tenailon et al. 2001), and a further reduction between public inbreds and private industry hybrids might be expected. Indeed, comparison of public and private industry inbreds (expired Plant Variety Protection) demonstrates limited genetic variation among many private inbreds, but also reveals some unique germplasm groups developed by private industry that were not represented in publicly developed lines (Nelson et al. 2008).

25.2 Breeding to Enhance Genetic Diversity in Elite Materials

The narrow genetic base of maize hybrids in the United States (all derived from only one of the 250 or so named races, the Corn Belt Dents) relative to the global diversity of the crop was recognized early on (Anderson 1944). Brown (1953) recommended the use of exotic germplasm to ameliorate the narrow genetic base of US maize and increase long-term potential for yield grain. Tropical maize, in particular, was identified as harboring the most genetic variation for observable characters, and as such the source of exotic germplasm most likely to have unique (and hopefully favorable) alleles absent from the Corn Belt Dents (Gerrish 1983; Goodman 1985; Tallury and Goodman 1999). In principal, the potential utility of broadening the genetic base of temperate maize is widely accepted, however the difficulties encountered by breeders in overcoming poor adaptation of tropical maize to temperate regions have hindered all efforts to broaden the genetic base of hybrid varieties grown in the USA (Hallauer 1978). Using tropical germplasm for breeding in temperate regions is hampered by: a lack of information needed to rationally choose exotic materials from among the tens of thousands of available sources; photoperiod sensitivity; a significant gap in agronomic quality between elite U.S. materials and exotic races; severe inbreeding depression; and undesirable agronomic characters such as weak roots and stalks, excessive plant and ear height, susceptibility to smut, and high grain moisture (Hallauer 1978; Goodman 1985, 1992).

Furthermore, the Corn Belt Dents have been generally recognized as one of the most, if not the most, inherently productive races of maize. It has been argued that their dominance in temperate regions is not by chance, but rather because they represent a hybrid race (admixture of Northern Flints and Southern Dents) and have the longest history of selection in the Corn Belt region of the USA (Troyer 1999). Recall that the Native American and Mesoamerican peoples that have the longest history as maize breeders did not grow maize in the grasslands that are now the Corn Belt region (Weatherwax 1954); rather, maize was not selected for these regions until the relatively recent transformation of the Midwestern prairies into farmland. Nevertheless, it is unlikely that the Corn Belt Dent race has a monopoly on all of the useful genes available in maize, and therefore, exotic germplasm may be useful for the improvement of U.S. maize (Brown 1975; Geadelmann 1984). Furthermore, as the dominant corn growing environments change due to global climate change and climatic events such as drought become more frequent, the utility of maize types selected by Native Americans for harsher environments may become essential.

Sources of exotic germplasm available to breeders include landrace accessions, composite populations, inbred lines, and hybrids. The breeding experience of Goodman et al. (2004) with tropical maize germplasm in the temperate USA has been a clear demonstration of the substantially greater utility of tropical hybrids and inbreds as breeding parents, as compared to recurrent selection populations, or, worse, landrace collections *per se*. The previous efforts of breeders in tropical regions in purging deleterious alleles during inbreeding to develop lines should be taken advantage of if at all possible. Starting with tropical hybrids, inbred lines with purely

tropical backgrounds but adapted to temperate regions have been developed by traditional breeding methods; although these lines have relatively poor performance *per se*, they produce hybrids with very high yield potential in some cases (Holley and Goodman 1988; Uhr and Goodman 1995a, b; Tallury and Goodman 1999; Goodman et al. 2000; Goodman 2004).

Breeding with landraces *per se* is more challenging, and the first difficulty is deciding which landrace accession to choose to use in a breeding program. Approximately 20,000 unique accessions of Latin American maize are stored in germplasm banks worldwide (Goodman 1983), and besides race name, there is often no information available to guide selection of starting materials. Evaluation of accessions for yield and agronomic performance *per se* in environments to which they are adapted is probably the most efficient and useful criterion for selection of breeding material. Castillo-Gonzalez and Goodman (1989) evaluated about 1,300 Latin American accessions in short daylength nurseries, and used their yield levels from this experiment as a culling criterion. The best accessions selected from this evaluation were crossed to a temperate line and selection within these breeding crosses resulted in the development of families and inbreds with acceptable adaptation and superior combining ability. (Holland and Goodman 1995; Tarter et al. 2003). Following this model, The Latin American Maize Project was undertaken to evaluate as many possible accessions in their home environments. Landrace collections were evaluated in their countries of origin (Salhuana et al. 1998) and the best collections were advanced to the Germplasm Enhancement of Maize (GEM) program (Pollak 2003). In the traditional GEM breeding protocol, superior landraces are crossed to elite proprietary inbreds, and the segregating populations are made available to GEM cooperators. Early generation selection is followed by extensive evaluations, and numerous lines with superior agronomic performance have been released for public use (Balint-Kurti et al. 2006). These programs have demonstrated the excellent potential of tropical maize germplasm for improving temperate material.

25.3 QTL Analysis and its Discontents

Many agriculturally and evolutionarily important traits in plants are quantitative in nature. Phenotypic variation for these traits is caused by a combination of segregation at multiple quantitative trait loci (QTL), the environment, and the interaction between genes and the environment (Mackay 2001). Two of the most commonly used approaches to dissect genes underlying complex quantitative traits are linkage analysis and association mapping (Mackay 2001; Risch and Merikangas 1996). Linkage analysis utilizes the shared inheritance of functional polymorphisms and adjacent markers within families or pedigrees of known ancestry. In plants, linkage analysis has been traditionally conducted with experimental populations derived from a biparental cross, such as F₂, backcross or recombinant inbred lines. Following the initial successes of identifying QTL in plants (Edwards et al. 1987; Paterson et al. 1988), methods to use QTL information to enhance selection of quantitative

traits were developed (Stuber and Edwards 1986; Lande and Thompson 1990). In particular, Tanksley and colleagues recognized the potential utility of linkage mapping approaches to aid the identification of unique favorable alleles in wild relatives and germplasm collections, and their subsequent incorporation into elite breeding populations (Tanksley and Nelson 1996; Tanksley and McCouch 1997). Indeed, the major practical success of QTL mapping has been the identification and marker-aided selection of QTL with moderate to large effects on biotic and abiotic stress resistances in several self-pollinating crops (Young 1999; Frary et al. 2000; Monforte and Tanksley 2000; Holland 2004; Pumphrey et al. 2007; Venuprasad et al. 2011).

Tuberosa and Salvi (2009) reviewed progress in QTL mapping in maize, citing several cases where QTL with moderate effects on complex traits such as abiotic stress resistance were identified, providing potential marker-assisted selection targets. In general this has been aided by the physiological dissection of complex traits into component traits (e.g., traits such as root architecture, leaf morphology, or anthesis-silk interval that influence grain yield), which have simpler genetic control when evaluated under controlled environmental conditions (Tuberosa and Salvi 2009). In general, however, most quantitative traits in maize appear to be under more complex genetic control relative to self-pollinating species. Thus, the genetic control is distributed across many loci, resulting in numerous loci with small effects, a situation in which QTL mapping has limited power and poor accuracy in typical mapping population sizes of a few hundred progeny lines (Beavis 1998; Melchinger et al. 1998). In such cases, very large population sizes are required to obtain accurate estimates of QTL positions and effects (Laurie et al. 2004; Schön et al. 2004; Holland 2007). Furthermore, the very high genetic diversity and low levels of linkage disequilibrium in diverse maize populations hinders the translation of QTL effect estimates from mapping populations to breeding populations representative of elite breeding programs (Holland 2004; Holland 2007; Bernardo 2008).

Recognizing the limited inferences that can be drawn from traditional biparental mapping populations and the difficulty in applying QTL mapping information to general breeding populations, maize geneticists pioneered methods to increase mapping resolution with advanced intercross line (AIL) populations and to broaden the inference space of QTL analyses by combining QTL mapping information across populations and pedigrees. The intermated B73 × Mo17 AIL population was derived by selfing lines to high levels of homozygosity following four generations of random mating, creating four times as many recombination events within small intervals compared to the initial F₂ generation (Lee et al. 2002; Sharopova et al. 2002; Winkler et al. 2003). This population serves as the community standard high resolution mapping population used to connect the B73 genome sequence and genetic map (Fu et al. 2006; Schnable et al. 2009), and also has been used for high resolution QTL mapping (Balint-Kurti et al. 2007; Lauter et al. 2008; Rodriguez et al. 2008; Zhang et al. 2010a). Other maize AILs have been used for genetic and QTL mapping (Falque et al. 2005; Falke et al. 2006; Huang et al. 2010b).

Meta-analysis of multiple independent QTL studies has been used to synthesize results with respect to a common consensus genetic map, highlighting genome regions that are consistently associated with variation for a trait across populations and

environments (Chardon et al. 2004) or improving the precision of QTL localization (Kump et al. 2010). A more direct approach to integrate QTL information across populations is joint population QTL mapping (Rebai et al. 1997; Blanc et al. 2006; Buckler et al. 2009; Coles et al. 2010). Joint linkage analysis increases power and resolution of QTL mapping, permits tests of QTL effect interactions with genetic backgrounds, and permits direct comparison of multiple allele effects, enhancing understanding of genetic heterogeneity (Holland 2007). Methods to combine information across more complex pedigrees have also been developed in the context of maize breeding programs (Zhang et al. 2005).

25.4 Association Analysis

Although linkage-based QTL mapping has been useful in identifying a number of genes affecting qualitative and quantitative traits, and despite substantial methodological advances pioneered in maize, several factors have hindered the translation of QTL mapping studies into breeding tools: the limitations of QTL mapping resolution (typically 10–20 cM, Holland 2007), accuracy of effect estimation, and sampling of allelic variation (typically only two alleles per locus, Holland 2004; Bernardo 2008). An alternative to linkage-based QTL mapping is association analysis, also known as association mapping or linkage disequilibrium mapping. Association analysis is based on gametic phase disequilibrium (commonly, although inaccurately, referred to as linkage disequilibrium, LD) to study the relationship between phenotypic variation and genetic polymorphisms. By focusing on diverse germplasm of unrelated ancestry, association analysis aims to sample genomes that have undergone thousands of generations of recombination since their descent from a common ancestor. As such, association mapping makes use of ancient as well as evolutionary recombination at the population level (Risch and Merikangas 1996; Remington et al. 2001; Thornsberry et al. 2001; Yu and Buckler 2006). The reduced correlations between even very closely linked loci potentially enables very high resolution marker-phenotype associations (Buckler and Thornsberry 2002; Flint-Garcia et al. 2005). Originally developed to identify genes involved in human diseases (Kerem et al. 1989; Corder et al. 1994), association mapping has become increasingly popular in plants in the last decade (Hauser et al. 2001; Thornsberry et al. 2001; Wilson et al. 2004; Szalma et al. 2005; Breseghello and Sorrells 2006a; Ehrenreich et al. 2009) because of advances in high throughput genomic technologies that provide dense coverage of the genome, the interest among breeders to identify novel and superior alleles, and improvements in statistical analysis methods. Advantages of association mapping over linkage mapping include the potential to survey effects of many alleles per locus, reduced cost and time to assemble an association mapping panel compared to creating structured populations for linkage analysis, and higher mapping resolution (Breseghello and Sorrells 2006a; Yu and Buckler 2006).

The resolution of association mapping is dependent upon the extent of linkage disequilibrium (LD), which, in turn, depends on recombination, genetic drift, selection, mating pattern and population admixtures; these factors vary both within species and between species (Flint-Garcia et al. 2003; Gaut and Long 2003). In maize, significant levels of LD extend less than 1 kb for landraces (Tenailon et al. 2001) and almost 2 kb for diverse inbred lines (Remington et al. 2001), but much farther in collections of elite commercial inbred lines (Ching et al. 2002; Rafalski 2002). Thus, in diverse maize samples, the rapid breakdown of LD is sufficient to permit gene-level mapping resolution, and is an ideal method to test the phenotypic effects of candidate genes, as has been done in maize for a handful of genes known or hypothesized to act as regulators or structural components of biochemical or developmental pathways. These include genes for flowering time, kernel composition, and secondary metabolite concentrations (Thornberry et al. 2001; Whitt et al. 2002; Palaisa et al. 2003; Wilson et al. 2004; Andersen et al. 2005; Szalma et al. 2005; Camus-Kulandaivelu et al. 2006; Harjes et al. 2008; Yan et al. 2010). Unfortunately, our understanding of genetic regulation is insufficient to reliably identify candidate genes for the vast majority of agriculturally important traits.

In the absence of a candidate gene list likely to contain causal loci, researchers must rely on random markers to sample the entire genome, in so-called genome-wide association studies (GWAS). The rate of decay of LD over physical distance determines the density of marker coverage needed to perform whole genome association analysis. In some self-pollinated crops or highly related maize populations with very extensive LD, one marker placed every cM or so can be sufficient to tag all segregating sites, but the resulting mapping resolution will be low (Brescghello and Sorrells 2006b; Rostoks et al. 2006; Hyten et al. 2007). In diverse maize, where LD decays rapidly, very high marker density is required to ensure a high probability that at least one marker is in high LD with causal loci (Yu and Buckler 2006). Gore et al. (2009) estimated that more than 10 million SNPs will be required to adequately conduct genome-wide association analysis in maize. The use of new high-throughput techniques, which allow genotyping hundreds of thousands of SNPs in a single assay and the creation of high density SNP haplotype maps in different plant species (Clark et al. 2007; Gore et al. 2009), has significantly boosted the application of association analysis in genome-wide scans for complex traits (Atwell et al. 2010; Brachi et al. 2010; Huang et al. 2010a; Kump et al. 2011; Poland et al. 2011; Ramsay et al. 2011; Tian et al. 2011). Association mapping undoubtedly has tremendous potential in dissecting the complex traits in plants and especially maize given its extensive phenotypic and molecular diversity. Several association mapping populations have been assembled in maize for various objectives (Andersen et al. 2005; Flint-Garcia et al. 2005; Camus-Kulandaivelu et al. 2006; Yan et al. 2009; Yang et al. 2010; Hansey et al. 2011).

Gene-phenotype associations may arise due to: causality (these are the associations we are most interested in), LD arising from physical proximity between marker and causal site (these can be useful in marker assisted selection), and LD arising from population structure. Population structure can cause highly significant associations between a marker and a phenotype, even when the marker is not physically linked to

any causative loci (Pritchard 2001; Thornsberry et al. 2001). Therefore, it is important to include estimates of population structure in the association analysis (Flint-Garcia et al. 2005). Various statistical approaches have been designed to control for population structure in different association samples such as the general linear model based approaches: genomic control (Devlin and Roeder 1999), and structured association (Pritchard et al. 2000) for population-based samples and transmission disequilibrium test (Abecasis et al. 2000) for family-based samples. Unified mixed linear model (MLM), and modified “compressed MLM” approaches appear to be superior to previously developed methods for association analysis in maize and other species (Yu et al. 2006; Zhang et al. 2010b). These methods can be used in the context of candidate gene association tests or GWAS (Zhang et al. 2010b). GWAS introduces computational challenges associated with conducting very large numbers of statistical tests, although increases in computer processing speed and improvements in algorithmic efficiency have permitted their application even with huge numbers of SNP tests in GWAS (Kang et al. 2008; Zhang et al. 2010b; Lippert et al. 2011). In addition, GWAS is confronted with the difficulty of determining significance thresholds for thousands or millions of statistical tests, although False Discovery Rate methods are very helpful in this regard as they are computationally tractable even with large numbers of tests (Benjamini and Yekutieli 2005).

25.5 Linkage and Association Analysis in Nested Association Mapping Populations

Linkage mapping and association analysis are complementary in many ways: linkage mapping has high power but low resolution, while association analysis has low power and high resolution. Linkage mapping uses structured populations to its advantage, while association analysis is hindered by population structure. To integrate the advantages of linkage analysis and association mapping into a single strategy, a large-scale set of inter-related maize mapping families was created to facilitate dissection of complex genetic variation underlying quantitative traits. The maize Nested Association Mapping (NAM) population was created to capture significant genetic variation and low linkage disequilibrium in a sample of lines that were used as founders to create multiple biparental linkage mapping populations (Yu et al. 2008; McMullen et al. 2009). The NAM population was created by crossing the inbred reference line B73 to 25 inbred lines that are representative of maize diversity (Flint-Garcia et al. 2005). From each cross, 200 recombinant inbred lines were derived by self-fertilization, resulting in a total of 5,000 RILs (McMullen et al. 2009). B73 was chosen as a reference line because of its role in the physical map and whole-genome sequence (Schnable et al. 2009). The other 25 parents maximize the diversity among the RIL families. More than half of these diverse lines are tropical in origin, nine are temperate lines, two are sweet corn lines and one is a popcorn line (McMullen et al. 2009). Thus, NAM is a specific case of inter-related mapping population mating designs referred to as a reference design. Although there are theoretically better mating designs for joint population QTL mapping (Verhoeven et al.

2006), the reference design was selected for practical reasons. Creating RILs with 50% pedigree contribution from a broadly-adapted temperate line ensured that even the effects of alleles from tropical inbred founders would not be highly confounded with poor adaptation to temperate environments. Each NAM line was genotyped with a common panel of 1,106 SNPs selected to cover the genome and to be polymorphic within most families, and the use of the common map permits investigation of recombination frequency differences among families and simplifies implementation of joint linkage QTL analysis (Buckler et al. 2009; McMullen et al. 2009).

The power of NAM for QTL analysis has been demonstrated by dissecting the genetic architecture of flowering time in maize. Joint linkage QTL mapping revealed 36–39 QTL affecting time to anthesis or silking (Buckler et al. 2009). All the identified QTL exerted small effect on the phenotype in an additive manner. In fact, the largest effect QTL for days to silking (DS) only had an additive effect of 1.7 days relative to B73. The complexity of flowering time in maize, in contrast to the identification of genes with very large effects on flowering time in self-pollinating species such as wheat and rice (Cockram et al. 2007; Izawa 2007), is likely due in part to the predominantly outcrossing mating system of maize. Rare mutations with large effects on flowering time would be associated with reproductive isolation and self-fertilization, and a consequent decrease in fitness in the progeny carrying the mutation. In addition, little evidence of epistasis or genotype-environment interactions (GEIs) was revealed, although the testing environments all had long daylengths; greater GEI would be expected when comparing across environments of distinct photoperiods (Buckler et al. 2009).

The most powerful application of NAM is the ability to efficiently conduct GWAS. NAM provides several substantial advantages for GWAS relative to diverse line association panels. First, the framework linkage map of 1,106 SNPs permits efficient imputation or “projection” of founder line SNP variation onto the entire RIL panel (Yu et al. 2008). For example, the maize haplotype map (HapMap version 1) consists of 1.6 million SNPs identified among the founders of the NAM population (Gore et al. 2009). Since the physical positions of those SNPs in the B73 reference sequence and their allelic composition among the founders are known, their probable allelic status in the NAM RILs can be imputed easily based on the flanking markers of the linkage map. Thus, the 1.6 M HapMap SNPs could be accurately imputed onto 5,000 mapping lines by sequencing only the 26 founders.

A second major advantage of GWAS in NAM is its known population structure. Whereas population structure in diverse line panels must be estimated with random markers, the structure of NAM is known: there are 25 families and there is no structure within families because the RILs within families were derived randomly. Thus, population structure is accounted for completely in the analysis simply by fitting the family main effect. Furthermore, the ability to conduct joint linkage QTL analysis and GWAS in the same population provides an additional advantage: the joint linkage QTL model can be used to account for genetic variation outside of the region being tested in GWAS, thus increasing the power for GWAS. The random derivation of RILs within each family also eliminates any unlinked LD that existed among the

founder lines, and dissipates LD among linked SNPs to an extent determined by the strength of linkage (Kump et al. 2011).

NAM-GWAS was conducted for leaf architecture traits (Tian et al. 2011) and two foliar diseases of maize (Kump et al. 2011; Poland et al. 2011). Similar to flowering time, the joint linkage QTL analysis detected between 29 to 36 loci for different traits (Poland et al. 2011; Tian et al. 2011). Again, most of the QTL effects were small, but together they explained more than 77–83 % of the genetic variance among RILs.

In these three studies, GWAS using 1.6 M HapMap SNPs identified between 203 and 295 SNPs with strong association with a trait. Among the associated SNPs, only 30–50 % of the SNPs for all three traits were in the QTL regions identified through linkage analysis. To some extent the incomplete overlap of QTL and associated SNP positions can be explained by low SNP coverage and differences in power of the two analyses. Some of the causative SNPs might have been missed in GWAS because the power to detect small effects that are segregating in only one or few crosses is limited (Holland 2007; Haley 2011). Also, complete marker saturation of the maize genome has been estimated to require ten times more SNPs than the current 1.6 M HapMap SNPs (Gore et al. 2009). Therefore it can be assumed that many of the causative SNPs were missed in the GWAS analyses due to a lack of linkage disequilibrium with the tested SNPs. Finally, only SNPs and small insertion/deletion polymorphisms were considered in these studies, whereas structural variation such as copy number variation and presence-absence variation among maize inbreds may also play a role in complex phenotypes (Lai et al. 2010; Swanson-Wagner et al. 2010; Eichten et al. 2011).

In spite of these limitations, each NAM-GWAS study (Kump et al. 2011; Poland et al. 2011; Tian et al. 2011) was able to successfully identify causal variation in or around several genes whose predicted functions are consistent with their association with the phenotype. The GWAS results provide candidate genes that can be tested further by fine-mapping and isolating individual loci affecting these important agronomic traits. However, one cannot be certain that SNPs identified as associated with a phenotypic trait in NAM-GWAS are in fact causal. In some cases, longer-range LD was observed between SNPs on the same chromosome because of the limited sampling of founders (Kump et al. 2011). It is likely that some proportion of SNPs associated with a trait in NAM-GWAS will turn out to be in LD with causal variants, perhaps at linked loci. The possibility that causal variants exist in non-coding regions (Clark et al. 2006; Salvi et al. 2007) and the lack of annotation for many predicted maize genes also hinders the interpretation of GWAS results in terms of biology.

Recently, the maize haplotype map has been expanded in terms of density of SNPs scored, types of variants scored (read depth variants, a proxy for copy number variation), and germplasm (now including more maize lines and some teosinte inbreds; Chia et al. 2012). This second generation maize haplotype map (HapMap II) provides more than 27 million SNPs and a thousand read depth variants scored on the NAM founders (www.panzea.org). This has provided more power for the most recent NAM-GWAS studies, but also complicates their interpretation, as separating false from true positive signals and causative variants from variants associated by LD becomes more difficult (Hung et al. 2012).

25.6 QTL Fine-Mapping

One way to proceed with identifying the causal variants associated with a QTL or a SNP associated with a phenotype is to conduct high-resolution fine mapping to resolve the variant to a single gene or single non-coding region. While QTL “cloning” is still a major challenge for most quantitative traits, it has been accomplished in a number of cases (Frery et al. 2000; Fridman et al. 2004; Salvi and Tuberosa 2005; Salvi et al. 2007). Selecting appropriate plant material before initiating a QTL fine-mapping and cloning experiment is perhaps the most important aspect of QTL characterization. Near-isogenic lines (NILs) and introgression lines (ILs) are often ideal material with which to initiate high-resolution mapping or positional cloning efforts (Fridman et al. 2004; Eichten et al. 2011). NILs generally refer to sets of lines differing from some common recurrent parent inbred by a small proportion of donor genome (in this case including the target QTL). Introgression lines (ILs) are backcross-derived lines containing segments from wild relatives (or exotic germplasm) such that an IL library would contain the entire genome of wild donor or exotic parent in the recurrent parent background (Zamir 2001; Salvi and Tuberosa 2005; Salvi et al. 2007).

Fine-mapping follows from QTL identification and development of a NIL pair differing only at the target QTL region. Markers defining the QTL region are used to select rare progeny with recombinant chromosomes from the segregating population derived from crossing the NIL pair (Fig. 25.2). Analysis of cosegregation between the phenotype and high density markers within the target region obtained from large insert genomic libraries (bacterial or yeast artificial chromosomes), a reference genome sequence where available, or next-generation sequencing technologies (Elshire et al. 2011) can resolve the QTL position to less than one cM genetic and less than Mb physical distances if sufficient recombinant progeny are obtained and marker density is high enough. With well-annotated sequence information on the narrowed QTL interval, candidate genes can be identified and their allelic sequence variants determined (Paran and Zamir 2003). Once candidate genes are identified, the final step is usually validating the phenotypic effect of critical sequence variants by genetic complementation tests, genetic engineering approaches such as RNA interference (RNAi), gene expression analyses, or by reverse genetic strategies like TILLING, T-DNA or transposon tagging. Several QTLs have been isolated in maize using these basic guidelines, e.g. *tb1* (Doebley et al. 1997; Doebley 2004), *tg1* (Wang et al. 2005), *DGAT1-2* (Zheng et al. 2008), *Rcg1* (Frey 2006; Frey et al. 2011), and *Vgt1* (Salvi et al. 2007).

Once an individual gene affecting a quantitative trait is isolated, the next step is to assess the molecular basis of allelic variation for that trait. In maize, variation in QTL alleles has been identified in both coding and regulatory regions of single genes. The cloned domestication QTL *tb1* (*teosinte branched1*) controls the difference in apical dominance between maize and its progenitor teosinte (Doebley et al. 1997), but no causal nucleotide differences have been observed in the *tb1* coding region between maize and teosinte alleles. Instead, the functional variation seems to lie in the regulatory region upstream of the *tb1* gene (Doebley et al. 1997; Wang et al. 1999; Clark

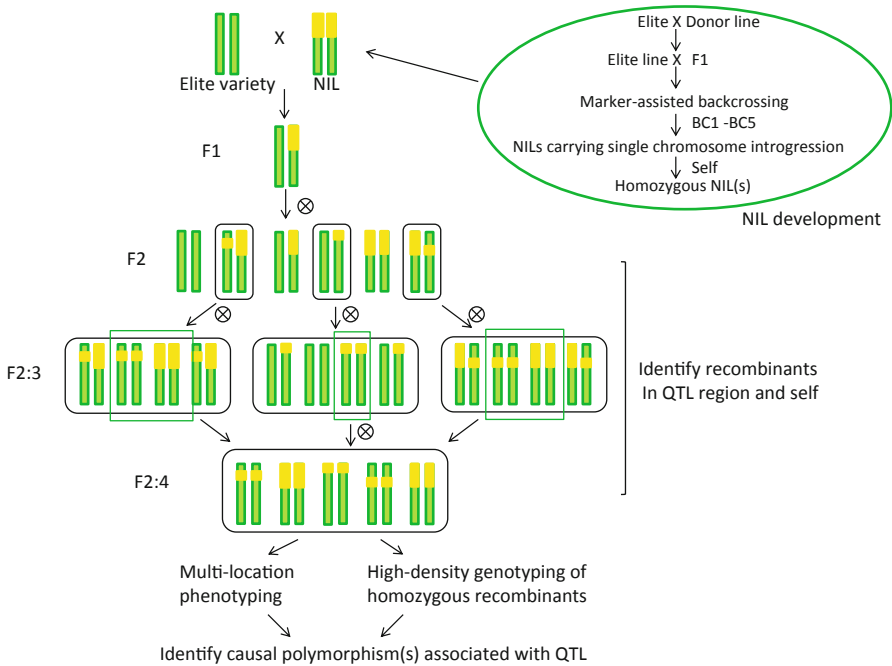


Fig. 25.2 Flow chart for high-resolution genetic mapping. Following the localization of a QTL to a $\sim 10\text{--}30\text{ cM}$ region of the genetic map, the causal gene(s) may be identified following homogenization of the genetic background and screening large segregating progenies for recombination events within the QTL interval. Multiple rounds of mapping and screening may be required to achieve gene-level resolution

et al. 2006). A transposon insertion in the regulatory region of *tb1* was shown to partially explain the increased apical dominance in maize compared to teosinte (Studer et al. 2011). Similarly, the flowering time QTL *Vgt1* (*vegetative to generative1*) is due to allelic variation in a noncoding region about 70 kb upstream of an *Ap2*-like transcription factor (Salvi et al. 2007). *Vgt1* acts as a cis-regulatory element that controls the expression of downstream genes (Salvi et al. 2007). In contrast, functional variation at the *tga1* (*teosinte glume architecture1*) locus which controls the differences in fruitcase/ear structure between maize and teosinte is due to the single amino acid substitution in the *tga1* protein (Wang et al. 2005). Similarly, *DGAT1-2*, a major QTL for high oil content, is caused by an amino acid insertion in the *DGAT1-2* protein in the ancestral allele that causes low oil content (Zheng et al. 2008). In addition to sequence variation leading to amino acid changes in proteins and altered regulatory activity, copy number variation (CNV) or presence absence variation (PAV) may underlie QTL. For example, *Rcg1* is a major QTL conferring resistance to anthracnose stalk rot, at which the resistant and susceptible alleles differ by the presence of an entire gene (Frey 2006; Frey et al. 2011). Only 5% of the US germplasm carries the resistant allele, but the allele is at higher frequency in tropical germplasm

(Frey 2006). The number of QTLs cloned so far is very small (Salvi and Tuberosa 2005; Doebley et al. 2006) and thus our knowledge about the genes and sequences causing such vast phenotypic variation is also very limited. Therefore, although difficult and costly, fine-mapping and cloning experiments provide unique information about the molecular basis of QTL.

Precise estimates of QTL positions are now available from joint linkage and GWAS analysis implemented in NAM (Buckler et al. 2009; Kump et al. 2011; Poland et al. 2011; Tian et al. 2011). We and others are attempting to identify the sequence variants (quantitative trait nucleotides, QTN) underlying some of these QTL by fine-mapping them in an effort to better understand the genetic control of complex traits and to validate and refine the statistical methods used for QTL mapping in NAM. In order to achieve detailed genetic characterization of QTLs underlying agronomic traits studied in NAM population, we have utilized a small sample of a series of near isogenic lines (NILs) carrying introgressions from the NAM founders in B73 genetic background developed by Syngenta AG (Pennisi 2008). We selected several NILs to target alleles with predicted significant effects on plant height, flowering time, and kernel composition as starting materials for fine-mapping. These NILs were crossed to B73 and segregating F₂ populations were screened with markers defining the introgression regions to identify recombinant progeny. Marker selection for homozygous selfed progenies of these F₂s was then used to obtain homozygous stocks carrying recombined-chromosomes in the target region (Fig. 25.2). With the availability of several million HapMap SNPs (www.panzea.org) and cost-effective genotyping platforms (including genotyping by sequencing, Elshire et al. 2011), it has become relatively efficient to densely genotype large number of genotypes in short time. However, a major challenge rests in accurately phenotyping the recombinant lines in the fine-mapping experiments, where the QTL has only a small phenotypic effect (Price 2006). Nonetheless, by replicating the phenotyping experiments within and across locations and by combining multiple data sets so as to increase the heritability of the trait (for low heritability traits), it might be possible to detect small QTL effects in fine-mapping populations. For now, it remains to be seen what sorts of genes are responsible for small effect variation in quantitative traits (or if the variants are coding regions at all), and the extent to which QTL identified in NAM were mapped precisely or how often QTL often represent statistical fusions of multiple linked genes that separate into distinct and possibly undetectable effects by high resolution mapping (Studer and Doebley 2011).

25.7 Marker-Based Selection for Complex Traits in Maize

Whereas marker-based selection has become routine for genes or QTL with moderate to large effects on agronomic traits in several self-pollinating species (Cahill and Schmidt 2004; Dubcovsky 2004; Collard and Mackill 2008; Jena and Mackill 2008; Yan et al. 2010), this has not generally been the case in maize to date, although there are a handful of specific traits where the substitution of markers for

phenotypic selection could become routine, e.g., kernel β -carotene content (Harjes et al. 2008; Yan et al. 2010) and anthracnose stalk rot resistance (Frey et al. 2011). Markers have been used to enhance phenotypic selection for quantitative traits like yield in maize (Crosbie et al. 2006; Eathington et al. 2007), but the implementation of marker-based QTL selection in maize has been hindered by the high diversity and low linkage disequilibrium in maize, both of which result in population-dependent marker-trait associations. In other words, QTL mapped in one biparental population may have little or no relation to the QTL segregating for the same trait in other breeding populations (Holland 2004). Efficient use of markers to enhance selection response for polygenic traits in maize requires identification of causal nucleotides (QTN) to use as reliable selection targets or breeding methods that more reliably relate genomic information to breeding values (Holland 2004; Bernardo 2008).

Although we are far from having lists of favorable agronomic trait QTN alleles from exotic germplasm, the very long-term goal of GWAS is to create such lists. Identifying QTN is most likely to occur first for component traits of very complex traits, because the components are more likely to be under simpler genetic control than the complex traits, as already demonstrated for QTL for yield components (Tuberosa and Salvi 2009). Once QTN are reliably known, diverse germplasm collections can be more effectively mined for unique allelic variants that may prove beneficial but are absent from elite breeding populations. Optimally, introgression libraries targeting a wide range of diversity at target regions could be used to estimate the allelic effects of the QTN. Developing nearly-isogenic stocks for each target gene from a wide range of germplasm is likely to be cost-prohibitive, however. If sequence information is available from very diverse germplasm collections, however, the sequence information can be used to selectively choose a small number of donors carrying the different variants at a gene. Unfortunately, given the very high level of sequence diversity in maize, it may generally be unclear which of the many sequence variants within or around a gene should be targeted, and there may be many haplotypes to test. An alternative strategy would be to evaluate earlier backcross generations, which will not as effectively isolate the QTN effect from the donor background, but are much easier and faster to develop. Coles et al. (2011) used this approach to validate several photoperiod response QTL and also to investigate the effects of QTL alleles derived from distinct backgrounds, revealing a surprising degree of variation in photoperiod response effect among tropical donor lines.

Another approach to characterize the effects of QTN across diverse germplasm sources would be to create synthetic populations that include contributions from very diverse germplasm sources, but which have been random-mated for a large number of generations to reduce linkage disequilibrium around most target genes, allowing high resolution association analysis of diverse allele effects to be conducted without the impediment of distinct alleles being nested within population subgroups, as can happen with association mapping in existing diverse germplasm panels. Finally, heterogeneous inbred families (HIFs) (Tuinstra et al. 1997) segregating at target regions in NAM or other mapping panels containing a diverse range of parents could be mined for near-isogenic pairs differing for alleles at that target region. HIFs are generally only available after mapping line development, however, and differences

in allelic contrasts among HIF pairs representing different parental combinations may be due to QTL—by—background interactions as well as allelic main effect differences.

If favorable QTN alleles can be identified, they can be selected for with diagnostic DNA marker assays in a straightforward manner, enabling breeders to more easily move them from exotic backgrounds to distantly related elite material. In the absence of knowledge of the sequence variants responsible for favorable QTL effects, however, selection for QTL alleles across unrelated populations is not expected to be effective because of genetic heterogeneity for complex traits (Holland 2007). To overcome the difficulty of genetic heterogeneity for QTL across breeding populations, industrial-scale breeding programs have implemented QTL mapping and selection within individual families. A typical breeding scheme might be as follows: (1) topcross doubled haploid (DH) lines from a breeding cross, (2) phenotype topcrosses in replicated trials in target production environment, (3) genotype the DH lines, (4) intermate selected DHs in year-round nursery, (5) repeat intermatings among individual progeny plants in year-round nurseries following seed or seedling selection for a desired marker profile, (6) create lines and topcrosses from final recurrent selection step, and (7) return lines and topcrosses to target production environment for re-evaluation. By mapping QTL for each family to be targeted for marker-based selection, the breeder is able to use QTL information directly relevant to the breeding family (Podlich et al. 2004). For example, Monsanto Co. has implemented marker-assisted recurrent selection in many breeding families, in which recurrent selection is conducted within each biparental family to increase the frequency of favorable QTL alleles mapped independently in each family (Crosbie et al. 2006; Eathington et al. 2007). This approach appears to be successful at enhancing genetic gains for yield above phenotypic selection, but the investment required in marker, breeding, winter nursery, statistical, and management infrastructure to use this form of MAS is very costly.

An alternative approach to implementing MAS within breeding families follows a similar breeding scheme, but uses genomic selection (GS) methods to create the marker-based selection index instead of QTL mapping to create the marker selection index, one can instead. Originally developed for animal breeding (Meuwissen et al. 2001), GS was introduced in the plant breeding literature in a maize breeding context similar to the one outlined above by Bernardo and Yu (2007). GS avoids the problem of distinguishing between false and true positive QTL, which underlies many of the statistical problems of QTL effect estimation and use for breeding (Beavis 1998; Schön et al. 2004), and instead fits all markers into a phenotype prediction model based on observed data. Obviously, this results in highly over-parameterized models, whose solution requires specialized statistical techniques which depend on assumptions made about the distribution of QTL effects (Bernardo and Yu 2007; Heffner et al. 2009; de los Campos et al. 2010; Lorenz et al. 2011). The distinguishing feature of all GS methods is that accurate estimation of *individual* marker effects is simply not a goal; instead, the objective is to obtain a model in which the *combined* effects of all marker loci provide accurate predictions of breeding values. GS generally provides greater response to selection than QTL-based marker-assisted recurrent selection when implemented within families (Bernardo and Yu 2007). Again, however,

the practical application of GS on a wide scale will require massive infrastructure to combine genotyping, off-season nursery management, statistical analysis, and very accurate seed and plant tracking.

Bernardo (2009) suggested that GS would be effective in exotic germplasm breeding programs in maize. He simulated GS in adapted—by—exotic populations with varying proportions of loci at which the exotic parent had the favorable allele. Phenotypic response to GS was predicted to be better in F_2 -derived populations than in backcross populations because of the higher probability of recovering favorable exotic alleles. Although GS appeared to reliably increase the frequency of favorable alleles from the adapted parent, the frequency of favorable alleles from the donor parent could decrease when the donor parent had a lower frequency of favorable alleles and when favorable exotic alleles were linked to unfavorable alleles (Bernardo 2009). Unfortunately, the reality is that we expect unfavorable linkage disequilibrium and relatively low frequencies of favorable exotic alleles to be the rule rather than the exception in germplasm incorporation programs.

An alternative approach to implementing GS is to attempt to build GS prediction models based on information from diverse breeding lines, such as those that might represent an entire breeding program (Heffner et al. 2009; Albrecht et al. 2011; Crossa et al. 2010; Lorenz et al. 2011) or even global maize diversity. Initial empirical tests of the predictive accuracy of GS models suggest that they should be effective for improving selection response among and within breeding crosses of elite lines (Albrecht et al. 2011; Crossa et al. 2010; Riedelsheimer et al. 2012). Prediction accuracy of breeding values for lines not closely related to the training populations is likely to be poor (Windhausen et al. 2012), however further research is needed to clarify the potential for GS in assisting breeding progress in adapted—by—exotic cross populations (Hamblin et al. 2011). A major difficulty that will continue to impede breeding by any method in such populations is the high proportion of linkages between unique favorable alleles from the exotic parent and alleles at nearby loci that cause problems in adaptation or poor agronomic performance. The crux of the problem is that GS methods are designed to select plants with highest predicted breeding value across their genome, but unadapted germplasm will almost always tend to have poor whole-genome breeding values even when it carries unique favorable alleles at a subset of loci. Thus, for GS to be effective at both increasing the mean genetic value of a breeding population and increasing the frequency of favorable alleles derived from exotic parents, it may need to be implemented in breeding populations that have undergone several to many generations of random-mating to break up linkages between favorable and unfavorable alleles derived from the exotic parent. Developing GS models in early generations of adapted—by—exotic cross populations may simply result in selection of progenies with higher proportions of adapted alleles.

In the following sections we outline three ongoing projects of which we are aware that are attempting to incorporate advanced genomics tools and strategies to increase the ability of maize breeders to identify favorable alleles and sources of germplasm for breeding.

25.8 GEM Allelic Diversity Project

The traditional GEM Project protocol involves selection among large numbers of early generation segregants from a limited number of exotic—by—adapted breeding families each year. Progress has been made in identifying lines with superior breeding values in these crosses, but this method restricts the number of breeding families that can be tested. In essence, the GEM project has been able to sample only a limited proportion of the favorable landrace accessions (representing a total of 24 races) identified by the LAMP project, and has not explored many other landraces deemed unacceptable by LAMP (Krakowsky et al. 2008). Furthermore, since each breeding cross involves crossing a landrace with one or two proprietary inbred lines, and the proprietary inbreds differ among crosses, GEM breeding crosses are not easily amenable to genomic analysis. Finally, lines released from the GEM program must meet minimal culling criteria for topcross yield potential and agronomic performance. For breeders interested in using exotic germplasm for specialty traits unrelated to yield, the potential elimination of lines carrying unique characteristics because of poor yield and agronomic performance may be undesirable. Therefore, in addition to the traditional breeding objective of the GEM project outlined above, a more recent effort organized by the GEM project has been to sample all of the Latin American races of maize via the “GEM Allelic Diversity Project”.

The Allelic Diversity project protocol involves crossing a sample of each race of the Latin American maize (about 250–300 collections in all) to each of two Pioneer Hybrid Corn Belt Dent inbred lines with expired Plant Variety Protection certificates. A small sample (3–5) of DH or selfed RILs from each cross will be created and propagated by self-fertilization if possible. The only selection criterion will be the line’s capacity to reproduce itself; otherwise, these will represent random, unselected lines from each cross. The resulting set of ~ 1,500 lines should represent the widest sample yet available of maize allelic diversity in common adapted genetic backgrounds and will serve as an excellent platform for allele mining (Krakowsky et al. 2008).

25.9 Seeds of Discovery—Large-scale Genotyping and Phenotyping of CIMMYT Germplasm

An initiative known as Seeds of Discovery funded by the Mexican Secretaria de Agricultura, Ganaderia, Desarrollo Rural, Pesca, y Alimentacion (SAGARPA) is being conducted collaboratively through CIMMYT with a number of Mexican Institutions. The objective of the Seeds of Discovery Initiative is to provide information on maize genetic resources that will facilitate their use by maize breeders in the developing world. A number of components are involved in this process: (1) genotyping all maize collections within Mexican Germplasm banks, (2) phenotypic characterization of Mexican germplasm bank accessions, (3) identification of haplotypes or alleles with effects on specific characteristics, (4) estimation of haplotype or allele

frequencies among and between representative accessions of races, (5) the creation of elite bridge lines carrying specific alleles or haplotype regions, and (6) formation of a web portal to deliver this information to breeders worldwide.

Priorities for phenotyping are related to climate change and delivering new alleles to breeders that will promote food security worldwide. These priorities are: drought and heat tolerance, resistance to diseases with expanding ranges due to climatic changes, and nitrogen and phosphorus use efficiency (Collins et al. 2008; Tuberosa 2012). Quality parameters for human consumption will also be included in the phenotypic characterization, in particular those characteristics most important for human consumption in México. Within México there is a strong desire to provide self-sufficiency in maize production. Half of the land devoted to maize production in México is planted to native landraces. To increase production in México, yields must increase both in areas devoted to improved maize cultivars (mostly hybrids) as well as those devoted to landraces. This can be accomplished by either displacing landraces with hybrids or by increasing yields of landraces, but the latter approach is preferred to maintain the genetic and cultural diversity of maize in México. A major objective of this project is to evaluate of the ability of landraces to produce a crop under suboptimal conditions while recognizing the specific culinary properties of diverse maize landraces. As the center of origin of maize, the continued use of landraces in México is important for the preservation of maize diversity as a world resource.

Phenotyping of a large GWAS experiment is currently underway. This initial experiment aims to identify favorable alleles for complex traits harbored in the 4,000 accessions of CIMMYT's breeders' core collection. One plant from each of the 4,000 accessions was used to pollinate a CIMMYT hybrid tester, and DNA was isolated from each landrace plant sampled. Genotyping by sequencing (Elshire et al. 2011) will be performed on the one plant per accession used as the male parent of each topcross. Topcross entries were assigned to sets of about 600 entries each according to the origin of the accession (lowland, subtropical or highland) to target entries to appropriate environments and to accommodate limitations of phenotyping capacity of collaborators. Although the design is not balanced, partial balance was achieved by including multiple sets (up to 2,200 topcross entries) in several environments, by including repeated check cultivars within and across environments, and by ensuring that at least 10 % of the accession topcross entries were planted at multiple locations. Testing locations used in the first season were a combination of Instituto Nacional de Investigaciones Forestales, Agrícolas Y Pecuaria (INIFAP) stations, INIFAP managed farmer's fields, Mexican University Field Stations, CIMMYT experiment stations, and one farmer's field managed by Pioneer Hybrid. The first season's evaluation will emphasize disease reaction, low nitrogen tolerance and agronomic traits. A follow-up evaluation in the second season will emphasize drought resistance, and quality for human consumption. This GWAS experiment will allow estimation of haplotype effects in topcrosses in target environments in México, serve as a training population for GS, and provide an estimate of how haplotype frequencies are distributed across germplasm groups.

Another component of the Seeds of Discovery project is to provide funding and coordinate projects proposed by Mexican institutions. In parallel with the topcross GWAS experiment, INIFAP is leading a large project to genotype and evaluate *per se* performance phenotypes of their most recent germplasm collection. These 6,000 were collected between 2008 and 2010 in México and are associated with accurate passport data, including GPS coordinates of collection sites. The goal is to produce *per se* phenotypic data that can be correlated with genotypic data at an allele frequency level. A first step is to determine through GIS data the most representative 1,200 accessions (20 %) to represent the agroclimatic diversity of México (Ruiz et al. 2008). The project is currently producing full-sib families from 30 plants per accession which have been individually sampled for DNA extraction. The hope is to estimate allele frequencies within 12 full sib families representing 24 individuals per accession. The 12 full sib families per accession will be phenotyped at multiple sites within their area of adaptation (e.g., tropical, subtropical, or highland environments). The phenotyping will be based on priorities within the adaptation zone. For example, drought tolerance and ear rot resistances are a priority for all target environments, but heat tolerance is a priority specifically for subtropical and tropical environments.

The collection will also be used to evaluate germplasm for specific culinary uses important in Mexican culture and food markets. Those uses which have the greatest added value for small farmers are a priority, such as pozole (hominy), elotes (sweet corn or green ears), and totomoxtle (husks for wrapping tamales). In addition, specific kernel quality characteristics for tortilla production will be evaluated.

Diversity of agroclimates and landraces make logistical considerations for phenotypic evaluation a challenge for this project. The GWAS experiment consists of maize accessions adapted to tropical, subtropical, and highland environments as well as temperate materials from South America. Phenotypic evaluation of these divergent materials cannot be conducted in a single common environment, and, further, the number of entries to be tested exceeds the phenotyping capacity of most cooperators. Therefore, the entries were partitioned into sets of materials based on agroclimate and maturity with 10 % overlapping entries in order to accommodate the logistical constraints of collaborators. The use of repeated commercial checks which have wide adaptation within México as well as use of repeated entries overlapping sites will permit combined statistical analysis of all environments. For specific disease hotspots we are using farmers' fields with high incidence of disease pressure. These fields are rainfed which presents specific constraints; rainfed conditions are particularly difficult to manage because planting occurs after the raining season starts, and the initiation of the rainy season in México has become unpredictable in recent years in México with rains starting a month or more later than historically expected. After the rains initiate, land must be prepared in a window of a few days when the fields are dry enough to enter with a tractor, but this is also unpredictable.

Additional logistical constraints in these very large-scale experiments, in addition to appropriate site selection and management, include sample tracking and environmental characterizations. Handheld data collection computers enabled with bar code readers will facilitate accurate data collection. Weather stations at each site and soil characterization will also be used to characterize the environmental factors related

to genotype-by-environment interactions. Investments for improving infrastructure at collaborator sites within México are underway to provide multiple sites with capability to conduct managed drought and heat screenings, as well as to improve seed storage capabilities and seed tracking management systems that will be necessary for the quantity of material to be evaluated during this initiative. Joint training of personnel between CIMMYT and Mexican Institutions is also an important part of this project, particularly the data management and bioinformatics portion of the project.

25.10 Bridging the Domestication Bottleneck with Teosinte Introgression Libraries

While maize inbreds and landraces contain an incredible amount of genetic diversity relative to other crop species, teosinte contains even more diversity than landraces and inbreds. Various population genetics studies indicate that maize inbreds retain approximately 60 % of the variation present in teosinte (Wright et al. 2005), and approximately 80 % of the variation present in landraces (Tenaillon et al. 2001). All genes across the genome experienced the domestication and/or breeding bottlenecks, resulting in moderate reductions in variation in maize relative to teosinte (Tenaillon et al. 2004). However, genes targeted by artificial selection during domestication and/or improvement have greatly reduced variation, as the combined effect of the bottleneck and selection is much more severe (Innan and Kim 2004). Thus teosinte should harbor more diversity for all genes compared to maize, and much more diversity for those genes that were targets of selection during domestication and/or plant breeding.

A population genetics study involving large-scale resequencing in maize revealed that 2–4 % of maize genes were targets of artificial selection during domestication and/or plant breeding (Wright et al. 2005). It is currently unknown what proportion of these selected genes were targets of selection during domestication (diversity lost between teosinte and landraces) versus improvement (diversity lost between landraces and inbred lines), or both. However, the implication of this study is striking. When considering the conservatively estimated filtered gene set of 32,690 genes (Schnable et al. 2009) or the more liberal estimate of 59,000 genes in maize (Messing et al. 2004), this implies that between 650 to 1,200 maize genes have experienced artificial selection, and have little or no sequence diversity in modern diverse inbreds, although they do in teosinte. Whereas these ~1,000 genes appear to have been under strong selection during the domestication and breeding of maize, this does not necessarily imply that the allele fixed in maize is the optimal allele for all modern environments and production systems. Furthermore, it is possible that suboptimal alleles were fixed in maize due to hitchhiking by tight linkage with a favorable allele at a nearby locus under selection (Tenaillon et al. 2002).

A large number of teosinte accessions can be obtained from either the USDA Plant Introduction Station in Ames, Iowa or the CIMMYT germplasm bank in México. However, direct comparison of maize to teosinte *per se* for any given trait is not

appropriate, as many of the undesirable teosinte traits (photoperiod sensitivity, incongruous plant architecture, lack of a true ear, the hard seed coat around the seed) mask potentially useful traits. Hence, teosinte must be crossed with maize to create germplasm that can be compared more equitably to maize. To this end, a set of introgression lines (ILs) is being developed from 10 *parviglumis* accessions in the B73 background.

Maize and *parviglumis* readily hybridize, both in the wild (Ellstrand et al. 2007) and in the nursery, given the proper conditions. As a short day plant, teosinte flowering is delayed under the long photoperiods of temperate US locations, and the first frost usually occurs prior to teosinte flowering. However, most teosintes can be induced to flower under short day conditions (Emerson 1924), and tassels can be observed in *parviglumis* within six weeks when grown in a day-neutral winter nursery site or growth chamber under short day conditions. When the objective is to make large numbers of crosses, it is easiest to conduct initial crosses involving teosinte in a winter nursery setting.

Using teosinte as the pollen parent in controlled pollinations is significantly easier than as the female parent for several reasons. Shoot-bagging teosinte is very difficult as silks often emerge from the axil prior to ear shoot appearance. There is also a potential for gametophyte factors to discriminate against or exclude pollen not carrying the same allele, although this is mostly a problem with the sister subspecies *Zea mays* ssp. *mexicana* (Kermicle and Allen 1990; Nelson 1994). Finally, a successful pollination using teosinte as the female would result in an ear with only 5–12 seeds, thus requiring significantly more work to generate large numbers of progeny.

The F₁ hybrids are sometimes still photoperiod sensitive, flowering around September 15 in Missouri and resulting in highly tillered plants with long lateral branches (although see Rogers 1950). Again, these symptoms appear to be alleviated in a short day environment. However, beginning with BC₁ plants, the process of backcrossing in a temperate environment becomes much easier (Fig. 25.3). The ultimate goal of the project is to produce BC₄ derived ILs, with the expected amount of teosinte being 3–5 % per line.

There are several ways that these introgression libraries will be used, and the applications described herein are interrelated. A very basic application is to explore empirical questions related to the processes of domestication and artificial selection. As described above, approximately 1,000 genes were targets of selection. Which genes are they and what are their functions? What traits were targeted by artificial selection during domestication/breeding? Are these selected genes relevant to agriculture today? An excellent example concerns a selected gene (AY104948) that has homology to the *Arabidopsis* *Auxin response factor1* (*ARF1*), a transcription factor with a putative function in plant growth. *ARF1* has very high levels of sequence diversity in teosinte, but almost no sequence diversity in maize inbreds (Wright et al. 2005). Auxin is clearly involved in apical dominance in plants, so it is possible that *ARF1* acts in a manner similar to *teosinte branched1* (Doebley et al. 1995). If so, we can postulate a corresponding phenotypic effect of the teosinte allele of *ARF1* in a maize background, such as increased tillering, increased lateral branching, and/or increased number of ears. However, preliminary studies of the teosinte introgression

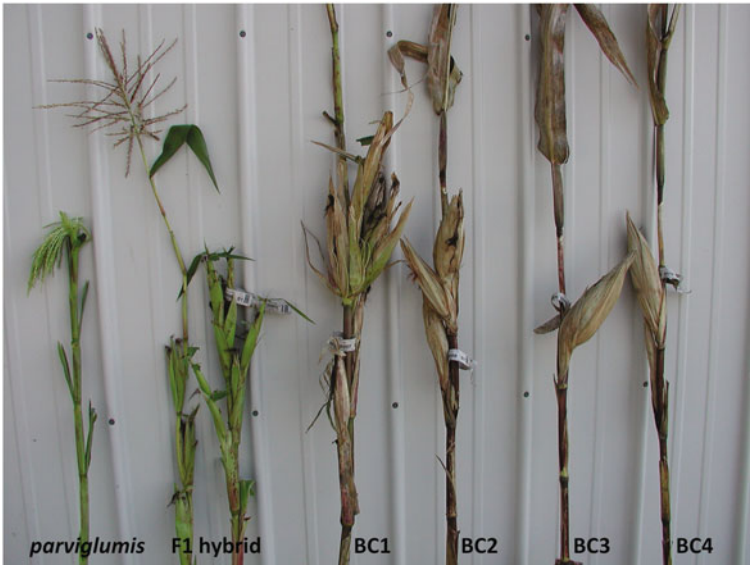


Fig. 25.3 Regression of progeny phenotypes to the B73 recurrent parent under repeated backcrossing of a maize-teosinte hybrid to the maize parent. Note, the ear heights shown do not represent the ear height on the plant. (Photographs by Sherry Flint-Garcia)

lines do not show an effect of *ARF1* on any of these traits. A comprehensive analysis of each of the $\sim 1,000$ selected genes is needed, and the teosinte introgression lines will play a vital role in testing hypotheses.

A second application is to evaluate and compare the range of allelic effects of teosinte to those of maize. Recent studies of the NAM population reveal that allele effect series are prevalent; in many cases a QTL segregates in multiple NAM families, but the direction and/or magnitude of the allelic effect varies across the NAM founders (Buckler et al. 2009; Kump et al. 2011; Poland et al. 2011; Tian et al. 2011). For example, for flowering time, variants at different sequence positions in *vgt1* result in opposite effects: a MITE insertion in *vgt1* is responsible for the early flowering Northern Flint allele (Salvi et al. 2007) and SNPs in its target gene, *rap2.7*, are likely responsible for a late flowering tropical allele (Buckler et al. 2009). Because *parviglumis* harbors many unselected, often deleterious, alleles that have not been purged by domestication and improvement, it will likely contain alleles with opposing effects as compared to maize. Furthermore, we postulate that a loss of genetic variation across the genome during domestication and/or breeding results in a loss of phenotypic variation, and therefore reintroduction of variation from teosinte will result in greater phenotypic variation. Following this logic, we hypothesize that teosinte harbors stronger alleles for any given QTL than maize. These stronger-effect teosinte alleles may be useful for genetic studies, such as in QTL fine mapping experiments as discussed above, or in physiological studies, where the objective is not necessarily to improve maize but rather to understand the genetic and/or physiological basis of complex traits.

A third application is the use of teosinte allelic variation for trait improvement. A more directed approach is to identify pathways controlling the trait of interest, and reintroduce variation from teosinte for genes involved in the pathway. For example, three genes in the starch pathway show signatures of past selection: the small subunit of ADP-glucose pyrophosphorylase encoded by *brittle2*, the starch branching enzyme encoded by *amylose extender1*, and the debranching enzyme encoded by *sugary1* (Whitt et al. 2002). Restoration of allelic variation from teosinte for these three selected genes could result in increased kernel starch content or alternate forms of starch that may be useful as specialty industrial starches and healthy, resistant (slow-degrading) starches. A second approach is more trait-focused, where the genes controlling the trait are perhaps unknown, but where teosinte shows greater trait variation than maize. For example, teosinte seeds contain twice the kernel protein content and novel zein proteins as compared to maize (Flint-Garcia et al. 2009a), as well as altered amino acid content (Flint-Garcia et al. 2009b). We hypothesize that variation from teosinte can be used to increase protein content and improve protein quality of maize. Indeed, an independent group of researchers has demonstrated that alien introgression lines of *Zea mays* ssp. *mexicana* have increased yield, protein content, and essential amino acid content compared to control lines (Wang et al. 2008a; Wang et al. 2008b).

Some have made the argument that the “best” alleles were already selected during domestication, and that reintroducing variation from teosinte would reverse human efforts over the last 9,000 years. In a few select cases this is true. Understandably, we do not want to reintroduce *tgal* alleles that confer the stony fruit case that surrounds the teosinte seed (Dorweiler et al. 1993). However, domestication occurred in a very different environment and under very different cultural practices than the USA Corn Belt. If maize were domesticated from teosinte in a temperate environment under modern agricultural practices then alternate alleles may well have been selected for many traits. We can capitalize on the incredible amount of diversity in teosinte to search for valuable alleles to aid in scientific discovery and continued corn improvement.

Acknowledgments Research by SS, SF-G, and JBH is supported by US National Science Foundation (DBI-0321467 and IOS-0820619). We thank Drs. Jesús Sánchez-Gonzalez (University of Guadalajara) and Major M. Goodman (North Carolina State University) for the ears and photographs used in Fig. 25.1.

References

- Abecasis G, Cardon L, Cookson W (2000) A general test of association for quantitative traits in nuclear families. *Am J Human Genet* 66:279–292
- Albrecht T, Wimmer V, Auinger H-J et al (2011) Genome-based prediction of testcross values in maize. *Theor Appl Genet* 123:339–350
- Andersen JR, Schrag T, Melchinger AE et al (2005) Validation of *Dwarf8* polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). *Theor Appl Genet* 111:206–217

- Anderson E (1944) The sources of effective germplasm in hybrid maize. *Ann MO Bot Gard* 31:355–361
- Anderson E, Cutler H (1942) Races of *Zea mays*: I. Their recognition and classification. *Ann MO Bot Gard* 29:69–88
- Atwell S, Huang YS, Vilhjalmsón BJ et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627–631
- Balint-Kurti PJ, Blanco M, Millard M et al (2006) Registration of 20 GEM maize breeding germplasm lines adapted to the southern USA. *Crop Sci* 46:996–998
- Balint-Kurti PJ, Zwonitzer JC, Wisser RJ et al (2007) Precise mapping of quantitative trait loci for resistance to southern leaf blight, caused by *Cochliobolus heterostrophus* race O, and flowering time using advanced intercross maize lines. *Genetics* 176:645–657
- Beavis WD (1998) QTL analyses: Power, precision, and accuracy. In: Paterson AH (ed) *Molecular dissection of complex traits*. CRC Press, Boca Raton, pp 145–162
- Benjamini Y, Yekutieli D (2005) Quantitative trait loci analysis using the false discovery rate. *Genetics* 171:783–789
- Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci* 48:1649–1664
- Bernardo R (2009) Genomewide selection for rapid introgression of exotic germplasm in maize. *Crop Sci* 49:419–425
- Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090
- Blanc G, Charcosset A, Mangin B et al (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theor Appl Genet* 113:206–224
- Brachi B, Faure N, Horton M et al (2010) Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet* 6:e1000940
- Breseghele F, Sorrells ME (2006a) Association analysis as a strategy for improvement of quantitative traits in plants. *Crop Sci* 46:1323–1330
- Breseghele F, Sorrells ME (2006b) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. *Genetics* 172:1165–1177
- Bretting PK, Goodman MM, Stuber CW (1987) Karyological and isozyme variation in West Indian and allied American mainland races of maize. *Am J Bot* 74:1601–1613
- Bretting PK, Goodman MM, Stuber CW (1990) Isozymatic variation in Guatemalan races of maize. *Am J Bot* 77:211–225
- Brown W (1953) Sources of germ plasm for hybrid corn. 8th Hybrid Corn Industry—Research Conference, pp 11–16
- Brown WL (1975) A broader germplasm base in corn and Sorghum. 30th Annual Corn and Sorghum Research Conference, pp 81–89
- Buckler ES, Holland JB, McMullen MM et al (2009) The genetic architecture of maize flowering time. *Science* 325:714
- Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. *Curr Opin Plant Biol* 5:107–111
- Cahill DJ, Schmidt DH (2004) Use of marker assisted selection in a product development breeding program. In: Fischer T, Turner N, Angus J, McIntyre L, Robertson M, Borrell A, Lloyd D (eds) *New directions for a diverse planet: Proc 4th Int Crop Sci Congress, Brisbane, Australia*
- Camus-Kulandaivelu L, Veyrieras JB, Madur D et al (2006) Maize adaptation to temperate climate: Relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* 172:2449–2463
- Castillo-Gonzalez F, Goodman MM (1989) Agronomic evaluation of Latin American maize accessions. *Crop Sci* 29:853–861
- Chardon F, Virlon B, Moreau L et al (2004) Genetic architecture of flowering time in maize as inferred from quantitative trait loci meta-analysis and synteny conservation with the rice genome. *Genetics* 168:2169–2185

- Chia JM, Song C, Bradbury PJ et al (2012) Maize hapmap 2 identifies extant variation from a genome in flux. *Nat Genet* 44:803–807
- Ching A, Caldwell KS, Jung M et al (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genet* 3
- Clark RM, Schweikert G, Toomajian C et al (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–342
- Clark RM, Wagler TN, Quijada P, Doebley J (2006) A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. *Nat Genet* 38:594–597
- Cockram J, Jones H, Leigh FJ et al (2007) Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity. *J Exp Bot* 58:1231–1244
- Coles ND, McMullen MD, Balint-Kurti PJ et al (2010) Genetic control of photoperiod sensitivity in maize revealed by joint multiple population analysis. *Genetics* 184:799–812
- Coles ND, Zila CT, Holland JB (2011) Allelic effect variation at key photoperiod response QTL in maize. *Crop Sci* 51:1036–1049
- Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos T Roy Soc B* 363:557–572
- Collins NC, Tardieu F, Tuberosa R (2008) QTL approaches for improving crop performance under abiotic stress conditions: where do we stand? *Plant Physiol* 147:469–486
- Corder EH, Saunders AM, Risch NJ et al (1994) Protective effect of apolipoprotein-E type-2 allele for late-onset Alzheimer disease. *Nat Genet* 7:180–184
- Crawford GW, Saunders D, Smith DG (2006) Pre-contact maize from Ontario, Canada: Context, chronology, variation, and plant association. In: Staller J, Tykot R, Benz B (eds) *Histories of maize: multidisciplinary approaches to the prehistory, linguistics, biogeography, domestication, and evolution of maize*. Academic Press, Burlington, pp 549–559
- Crosbie TM, Eathington SR, Johnson GR et al (2006) Plant breeding: past, present, and future. In: Lamkey KR, Lee M (eds) *Plant breeding: The Arnel R Hallauer International Symposium*. Blackwell, Ames, pp 3–50
- Crosby A (1972) *The Columbian exchange: biological and cultural consequences of 1492*. Greenwood, Westport, CT
- Crossa J, de los Campos G, Perez P et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–U406
- de los Campos G, Gianola D, Rosa GJM et al (2010) Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genet Res* 92:295–308
- Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 5:997–1004
- Doebley J (2004) The genetics of maize evolution. *Ann Rev Genet* 38:37–59
- Doebley J, Gaut BS, Smith BD (2006) The molecular genetics of crop domestication. *Cell* 127:1309–1321
- Doebley J, Stec A, Gustus C (1995) *teosinte branched* and the origin of maize: Evidence for epistasis and the evolution of dominance. *Genetics* 141:333–346
- Doebley J, Stec A, Hubbard L (1997) The evolution of apical dominance in maize. *Nature* 386:485–488
- Doebley J, Wendel JD, Smith JSC et al (1988) The origin of Cornbelt maize: the isozyme evidence. *Econ Bot* 42:120–131
- Doebley JF, Goodman MM, Stuber CW (1984) Isoenzymatic variation in *Zea (gramineae)*. *Syst Bot* 9:204–218
- Doebley JF, Goodman MM, Stuber CW (1985) Isozyme variation in the races of maize from Mexico. *Am J Bot* 72:629–639
- Dorweiler J, Stec A, Kermicle J, Doebley J (1993) *Teosinte-Glume-Architecture-1* – a genetic locus controlling a key step in maize evolution. *Science* 262:233–235
- Dubcovsky J (2004) Marker-assisted selection in public breeding programs: The wheat experience. *Crop Sci* 44:1895–1898

- Dubreuil P, Warburton M, Chastanet M et al (2006) More on the introduction of temperate maize into Europe: Large-scale bulk SSR genotyping and new historical elements. *Maydica* 51:281–291
- Duvick DN, Smith JSC, Cooper M (2004) Changes in performance, parentage, and genetic diversity of successful corn hybrids, 1930–2000. In: Smith CW, Betran FJ, Runge ECA (eds) *Corn: origin, history, technology, and production*. Wiley, New York, pp 65–97
- Eathington SR, Crosbie TM, Edwards MD et al (2007) Molecular markers in a commercial breeding program. *Crop Sci* 47:S-154–163
- Edwards MD, Stuber CW, Wendel JF (1987) Molecular-marker-facilitated investigations of quantitative-trait loci in maize. I. Numbers, genomic distribution, and types of gene action. *Genetics* 116:113–125
- Ehrenreich IM, Hanzawa Y, Chou L et al (2009) Candidate gene association mapping of Arabidopsis flowering time. *Genetics* 183:325–335
- Eichten SR, Foerster JM, de Leon N et al (2011) B73-Mo17 near-isogenic lines demonstrate dispersed structural variation in maize. *Plant Physiol* 156:1679–1690
- Ellstrand NC, Garner LC, Hegde S et al (2007) Spontaneous hybridization between maize and teosinte. *J Hered* 98:183–187
- Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple Genotyping-by-Sequencing (GBS) approach for high diversity species. *PLoS One* 6:e19379
- Emerson RA (1924) Control of flowering in teosinte. Short-day treatment brings early flowers. *J. Hered.* 15:41–48
- Falke KC, Melchinger AE, Flachenecker C et al (2006) Comparison of linkage maps from F2 and three times intermated generations in two populations of European flint maize (*Zea mays* L.). *Theor Appl Genet* 113:857–866
- Falque M, Decousset L, Dervins D et al (2005) Linkage mapping of 1454 new maize candidate gene loci. *Genetics* 170:1957–1966
- Flint-Garcia SA, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol* 54:357–374
- Flint-Garcia SA, Thuillet AC, Yu J et al (2005) Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant Journal* 44:1054–1064
- Flint-Garcia SA, Bodnar AL, Scott MP (2009a) Wide variability in kernel composition, seed characteristics, and zein profiles among diverse maize inbreds, landraces, and teosinte. *Theor Appl Genet* 119:1129–1142
- Flint-Garcia SA, Guill KE, Sanchez-Villeda H et al (2009b) Maize amino acid pathways maintain high levels of genetic diversity. *Maydica* 54:375–386
- Frary A, Nesbitt TC, Frary A et al (2000) *fw2.2*: A quantitative trait locus key to the evolution of tomato fruit size. *Science* 289:85–87
- Frey TJ (2006) Fine mapping, cloning, verification, and fitness evaluation of a QTL, *Rcg1*, which confers resistance to *Colletotrichum graminicola* in maize. Ph.D. Thesis. Dep Plant and Soil Sciences. Univ. Delaware, Newark, DE
- Frey TJ, Weldekidan T, Colbert T et al (2011) Fitness evaluation of *Rcg1*, a locus that confers resistance to *Colletotrichum graminicola* (Ces.) GW Wils. using near-isogenic maize hybrids. *Crop Sci* 51:1551–1563
- Fridman E, Carrari F, Liu Y-S et al (2004) Zooming in on a quantitative trait for the tomato yield using interspecific introgressions. *Science* 305:1786–1789
- Fu Y, Wen TJ, Ronin YI et al (2006) Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. *Genetics* 174:1671–1683
- Gaut BS, Long AD (2003) The lowdown on linkage disequilibrium. *Plant Cell* 15:1502–1506
- Geadelmann JL (1984) Using exotic germplasm to improve northern corn. 39th Annual Corn & Sorghum Research Conference, pp 98–110
- Gerrish EE (1983) Indications from a diallel study for interracial maize hybridization in the Corn Belt [Central USA]. *Crop Sci* 23:1082–1084

- Goodman MM (1983) Racial diversity in maize. In: Williams LE, Gordon DT, Nault LR (eds) International Maize Virus Disease Colloquium and Workshop. Ohio Agricultural Research and Development Center, Wooster, pp 29–40
- Goodman MM (1985) Exotic maize germplasm: Status, prospects, and remedies. *Iowa State J Res* 59:497–527
- Goodman MM (1992) Choosing and using tropical corn germplasm. 47th Annual Corn & Sorghum Research Conference. Am. Seed Trade Assoc., Washington, DC, pp 47–64
- Goodman MM (2004) Developing temperate inbreds using tropical maize germplasm: Rationale, results, conclusions. *Maydica* 49:209–219
- Goodman MM, Brown WL (1988) Races of corn. In: Sprague GF, Dudley JW (eds) Corn and corn improvement. Am Soc Agron, Madison, pp 33–79
- Goodman MM, Moreno J, Castillo F et al (2000) Using tropical maize germplasm for temperate breeding. *Maydica* 45:221–234
- Goodman MM, Stuber CW (1983) Races of maize. VI. Isozyme variation among races of maize in Bolivia [*Zea mays*, corn]. *Maydica* 28:169–187
- Gore MA, Chia JM, Elshire RJ et al (2009) A first-generation haplotype map of maize. *Science* 326:1115–1117
- Haley C (2011) A cornucopia of maize genes. *Nat Genet* 43:87–88
- Hallauer AR (1978) Potential of exotic germplasm for maize improvement. In: Walden DB (ed) Maize breeding and genetics. Wiley, New York, pp 229–247
- Hamblin MT, Buckler ES, Jannink J-L (2011) Population genetics of genomics-based crop improvement methods. *Trends Genet* 27:98–106
- Hansey CN, Johnson JM, Sekhon RS et al (2011) Genetic diversity of a maize association population with restricted phenology. *Crop Sci* 51:704–715
- Harjes CE, Rocheford TR, Bai L et al (2008) Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* 319:330–333
- Hauser MT, Harr B, Schlotterer C (2001) Trichome distribution in *Arabidopsis thaliana* and its close relative *Arabidopsis lyrata*: Molecular analysis of the candidate gene *GLABROUS1*. *Mol Biol Evol* 18:1754–1763
- Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
- Hernández E (1985) Maize and man in the Greater Southwest. *Econ Bot* 39:416–430
- Holland JB (2004) Implementation of molecular markers for quantitative traits in breeding programs—challenges and opportunities. In: Fischer T, Turner N, Angus J, McIntyre L, Robertson M, Borrell A, Lloyd D (eds) New directions for a diverse planet: Proc 4th Int Crop Sci Congress, Brisbane, Australia
- Holland JB (2007) Genetic architecture of complex traits in plants. *Curr Opin Plant Biol* 10:156–161
- Holland JB, Goodman MM (1995) Combining ability of tropical maize accessions with U.S. germplasm. *Crop Sci* 35:767–773
- Holland JB, Nelson PT (2010) Dedication: Major M. Goodman: Maize Geneticist and Breeder. *Plant Breed Rev*. Wiley, pp 1–29
- Holley RN, Goodman MM (1988) Yield potential of tropical hybrid maize derivatives. *Crop Sci* 28:213–218
- Huang X, Wei X, Sang T et al (2010a) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–967
- Huang Y-F, Madur D, Combes V et al (2010b) The genetic architecture of grain yield and related traits in *Zea mays* L. revealed by comparing intermated and conventional populations. *Genetics* 186:395–404
- Hung HY, Shannon LM, Tian F et al (2012) ZmCCT and the genetic basis of day-length adaptation underlying the post-domestication spread of maize. *Proc Natl Acad U S A* 109:E1913–1921
- Hyten DL, Choi IY, Song QJ et al (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175:1937–1944

- Innan H, Kim Y (2004) Pattern of polymorphism after strong artificial selection in a domestication event. *Proc Nat Acad U S A* 101:10667–10672
- Izawa T (2007) Adaptation of flowering-time by natural and artificial selection in Arabidopsis and rice. *J Exp Bot* 58:3091–3097
- Jena KK, Mackill DJ (2008) Molecular markers and their use in marker-assisted selection in rice. *Crop Sci* 48:1266–1276
- Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–1723
- Kerem BS, Rommens JM, Buchanan JA et al (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science* 245:1073–1080
- Kermicle JL, Allen JO (1990) Cross-incompatibility between maize and teosinte. *Maydica* 35:399–408
- Krakowsky MD, Holley R, Deutsch JA et al (2008) Maize allelic diversity project. 50th Maize Genetics Conference, Washington, DC
- Kuleshov NN (1933) World diversity of phenotypes of maize. *J Amer Soc Agron* 25:688–700
- Kump KL, Bradbury PJ, Buckler ES et al (2011) Genome-wide association study of quantitative resistance to Southern leaf blight in the maize nested association mapping population. *Nat Genet* 43:163–168
- Kump KL, Holland JB, Jung MT et al (2010) Joint analysis of near-isogenic and recombinant inbred line populations yields precise positional estimates for quantitative trait loci. *Plant Genome* 3:142–153
- Lai JS, Li RQ, Xu X et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. *Nat Genet* 42:1027–U1158
- Lande R, Thompson R (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124:743–756
- Laurie CC, Chasalow SD, Ledeaux JR et al (2004) The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. *Genetics* 168:2141–2155
- Lauter N, Moscou MJ, Habiger J, Moose SP (2008) Quantitative genetic dissection of shoot architecture traits in maize: towards a functional genomics approach. *Plant Genome* 1:99–110
- Lee M, Sharopova N, Beavis WD et al (2002) Expanding the genetic map of maize with the intermated B73 x Mo17 (*IBM*) population. *Plant Mol Biol* 48:453–461
- Lippert C, Listgarten J, Liu Y et al (2011) FaST linear mixed models for genome-wide association studies. *Nat Meth* 8:833–835
- Lorenz AJ, Chao SM, Asoro FG et al (2011) Genomic selection in plant breeding: Knowledge and prospects. *Adv Agron* 110:77–123
- Mackay TFC (2001) The genetic architecture of quantitative traits. *Ann Rev Genet* 35:303–309
- Mangelsdorf PC (1974) Corn: its origin, evolution, and improvement. Belknap Press of Harvard University Press, Cambridge
- Matsuoka Y, Vigouroux Y, Goodman MM et al (2002) A single domestication for maize shown by multilocus microsatellite genotyping. *Proc Nat Acad U S A* 99:6080–6084
- McMullen MD, Kresovich S, Sanchez Villeda H et al (2009) Genetic properties of the maize Nested Association Mapping population. *Science* 325:737–740
- Melchinger AE, Utz HF, Schön CC (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveal low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149:383–403
- Messing J, Bharti AK, Karlowski WM et al (2004) Sequence composition and genome organization of maize. *Proc Nat Acad U S A* 101:14349–14354
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Monforte AJ, Tanksley SD (2000) Fine mapping of a quantitative trait locus (QTL) from *Lycopersicon hirsutum* chromosome 1 affecting fruit characteristics and agronomic traits: breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. *Theor Appl Genet* 100:471–479

- Nelson OE (1994) The gametophyte factors of maize. In: Freeling M, Walbot V (eds) *The maize handbook*. Springer-Verlag, New York, pp 496–503
- Nelson PT, Coles ND, Holland JB et al (2008) Molecular characterization of maize inbreds with expired U.S. plant variety protection. *Crop Sci* 48:1673–1685
- Palaisa KA, Morgante M, Williams M, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell* 15:1795–1806
- Paran I, Zamir D (2003) Quantitative traits in plants: beyond the QTL. *Trends Genet* 19:303–306
- Paterson AH, Lander ES, Hewitt JD et al (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* 335:721–726
- Pennisi E (2008) Plant sciences—Corn genomics pops wide open. *Science* 319:1333–1333
- Piperno DR, Ranere AJ, Holst I et al (2009) Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, México. *Proc Natl Acad Sci U S A* 106:5019–5024
- Podlich DW, Winkler CR, Cooper M (2004) Mapping as you go. An effective approach for marker-assisted selection of complex traits. *Crop Sci* 44:1560–1571
- Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proc Nat Acad U S A* 108:6893–6898
- Pollak LM (2003) The history and success of the public-private project on germplasm enhancement of maize (GEM). *Adv Agron* 78:45–87
- Pressoir G, Berthaud J (2004a) Patterns of population structure in maize landraces from the Central Valleys of Oaxaca in Mexico. *Heredity* 92:88–94
- Pressoir G, Berthaud J (2004b) Population structure and strong divergent selection shape phenotypic diversification in maize landraces. *Heredity* 92:95–101
- Price AH (2006) Believe it or not, QTLs are accurate!. *Trends Plant Sci* 11:213–216
- Pritchard JK (2001) Deconstructing maize population structure. *Nat Genet* 28:203–204
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics* 155:945–959
- Pumphrey MO, Bernardo R, Anderson JA (2007) Validating the QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. *Crop Sci* 47:200–206
- Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin Plant Biol* 5:94–100
- Ramsay L, Comadran J, Druka A et al (2011) *INTERMEDIUM-C*, a modifier of lateral spikelet fertility in barley, is an ortholog of the maize domestication gene *TEOSINTE BRANCHED 1*. *Nat Genet* 43:169–172
- Rebai A, Blanchard P, Perret D, Vincourt P (1997) Mapping quantitative trait loci controlling silking date in a diallel cross among four lines of maize. *Theor Appl Genet* 95:451–459
- Rebourg C, Chastanet M, Gouesnard B et al (2003) Maize introduction into Europe: the history reviewed in the light of molecular data. *Theor Appl Genet* 106:895–903
- Reif J, Warburton M, Xia X et al (2006) Grouping of accessions of Mexican races of maize revisited with SSR markers. *Theor Appl Genet* 113:177–185
- Remington DL, Thornsberry JM, Matsuoka Y et al (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Nat Acad U S A* 98:11479–11484
- Riedelsheimer C, Czedik-Eysenberg A, Grieder C et al (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* 44:217–220
- Rincón SF, Castillo GF, Ruiz T NA (2010) Diversidad y distribución de los maíces nativos en Coahuila, México. *SOMEFI*, Chapingo, México
- Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273:1516–1517
- Rodríguez VM, Butron A, Malvar RA et al (2008) Quantitative trait loci for cold tolerance in the maize IBM population. *Int J Plant Sci* 169:551–556

- Rogers JS (1950) The inheritance of photoperiodic response and tillering in maize-teosinte hybrids. *Genetics* 35:513–540
- Ron Parra J, Sánchez-González JJ, Jiménez-Cordero AA et al (2006) Maíces nativos del Occidente de México I. *Colectas* 2004. *Scientia-CUCBA* 8:1–139
- Rostoks N, Ramsay L, MacKenzie K et al (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Nat Acad U S A* 103:18656–18661
- Ruiz C JA, Puga ND, Sánchez G JJ et al (2008) Climatic adaptation and ecological descriptors of 42 Mexican maize races. *Crop Sci* 48:1502–1512
- Salhuana W, Pollak LM, Ferrer M et al (1998) Breeding potential of maize accessions from Argentina, Chile, USA, and Uruguay. *Crop Sci* 38:866–872
- Salvi S, Sponza G, Morgante M et al (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc Nat Acad U S A* 104:11376–11381
- Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. *Trends Plant Sci* 10:297–304
- Sanchez G JJ, Goodman MM (1992a) Relationships among Mexican and some North American and South American races of maize. *Maydica* 37:41–51
- Sanchez G JJ, Goodman MM (1992b) Relationships among the Mexican races of maize. *Econ Bot* 46:72–85
- Sanchez G JJ, Goodman MM, Stuber CW (2000a) Isozymatic and morphological diversity in the races of maize of Mexico. *Econ Bot* 54:43–59
- Sanchez G JJ, Stuber CW, Goodman MM (2000b) Isozymatic diversity in the races of maize of the Americas. *Maydica* 45:185–203
- Sanchez G JJ, Goodman MM, Bird RMK, Stuber CW (2006) Isozyme and morphological variation in maize of five Andean countries. *Maydica* 51:25–42
- Sanchez G JJ, Goodman MM, Stuber CW (2007) Racial diversity of maize in Brazil and adjacent areas. *Maydica* 52:13–30
- Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: Complexity, diversity, and dynamics. *Science* 326:1112–1115
- Schön CC, Utz HF, Groh S et al (2004) Quantitative trait locus mapping based on resampling in a vast maize testcross experiment and its relevance to quantitative genetics for complex traits. *Genetics* 167:485–498
- Sharopova N, McMullen MD, Schultz L et al (2002) Development and mapping of SSR markers for maize. *Plant Mol Biol* 48:463–481
- Smith BD (1989) Origins of agriculture in eastern North America. *Science* 246:1566–1571
- Smith JSC, Smith OS, Wright S et al (1992) Diversity of U.S. hybrid maize germplasm as revealed by restriction fragment length polymorphisms. *Crop Sci* 32:598–604
- Stuber CW, Edwards MD (1986) Genotypic selection for improvement of quantitative traits in corn using molecular marker loci. *Proc 41st Ann Corn & Sorghum Res Conf. American Seed Trade Association, Chicago, IL*, pp 70–83
- Studer A, Zhao Q, Ross-Ibarra J, Doebley J (2011) Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat Genet* 43:1160–1163
- Studer AJ, Doebley JF (2011) Do large effect QTL fractionate? A case study at the maize domestication QTL *teosinte branched1*. *Genetics* 188:673–681
- Sturtevant E (1899) Varieties of corn. *US Dep Agr Off Exp Sta Bul* 57
- Swanson-Wagner RA, Eichten SR, Kumari S et al (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res* 20:1689–1699
- Szalma S, Buckler E, Snook M, McMullen M (2005) Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks. *Theor Appl Genet* 110:1324–1333
- Tallury SP, Goodman MM (1999) Experimental evaluation of the potential of tropical germplasm for temperate maize improvement. *Theor Appl Genet* 98:54–61
- Tanksley SD, McCouch SR (1997) Seed banks and molecular maps: Unlocking genetic potential from the wild. *Science* 277:1063–1066

- Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor Appl Genet* 92:191–203
- Tarter JA, Goodman MM, Holland JB (2003) Testcross performance of semiexotic inbred lines derived from Latin American maize accessions. *Crop Sci* 43:2272–2278
- Tenaillon MI, Sawkins MC, Long AD et al (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Nat Acad U S A* 98:9161–9166
- Tenaillon MI, Sawkins MC, Anderson LK et al (2002) Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Genetics* 162:1401–1413
- Tenaillon MI, U'Ren J, Tenaillon O, Gaut BS (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. *Mol Biol Evol* 21:1214–1225
- Thornsberry JM, Goodman MM, Doebley J et al (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 28:286–289
- Tian F, Bradbury PJ, Brown PJ et al (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* 43:159–162
- Troyer AF (1999) Background of U.S. hybrid corn. *Crop Sci* 39:601–626
- Tuberosa R (2012) Phenotyping for drought tolerance of crops in the genomics era. *Frontiers in Plant Physiol* 3(347):1–25
- Tuberosa R, Salvi S (2009) QTL for agronomic traits in maize production. In: Bennetzen JL, Hake SC (eds) *Handbook of maize: its biology*. Springer, New York, pp 501–541
- Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. *Theor Appl Genet* 95:1005–1011
- Uhr DV, Goodman MM (1995a) Temperate maize inbreds derived from tropical germplasm: I. Testcross yield trials. *Crop Sci* 35:779–784
- Uhr DV, Goodman MM (1995b) Temperate maize inbreds derived from tropical germplasm: II. Inbred yield trials. *Crop Sci* 35:785–790
- Van Heerwaarden J, Doebley J, Briggs WH et al (2011) Genetic signals of origin, spread, and introgression in a large sample of maize landraces. *Proc Natl Acad Sci USA* 108:1088–1092
- Venuprasad R, Bool M, Quiatchon L, Atlin G (2011) A QTL for rice grain yield in aerobic environments with large effects in three genetic backgrounds. *Theor Appl Genet* 1–10
- Verhoeven KJF, Jannink JL, McIntyre LM (2006) Using mating designs to uncover QTL and the genetic architecture of complex traits. *Heredity* 96:139–149
- Vigouroux Y, Glaubitz JC, Matsuoka Y et al (2008) Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. *Am J Bot* 95:1240–1253
- Wang H, Nussbaum-Wagler T, Li BL et al (2005) The origin of the naked grains of maize. *Nature* 436:714–719
- Wang LZ, Xu CZ, Qu ML, Zhang JR (2008a) Kernel amino acid composition and protein content of introgression lines from *Zea mays* ssp. *mexicana* into cultivated maize. *J Cereal Sci* 48:387–393
- Wang LZ, Yang AF, He CM et al (2008b) Creation of new maize germplasm using alien introgression from *Zea mays* ssp. *mexicana*. *Euphytica* 164:789–801
- Wang R-L, Stec A, Hey J et al (1999) The limits of selection during maize domestication. *Nature* 398:236–239
- Weatherwax P (1954) *Indian corn in old America*. McMillan, New York
- Whitt SR, Wilson LM, Tenaillon MI et al (2002) Genetic diversity and selection in the maize starch pathway. *Proc Nat Acad U S A* 99:12959–12962
- Wilson LM, Whitt SR, Ibañez AM et al (2004) Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* 16:2719–2733
- Windhausen VS, Atlin GN, Hickey JM et al (2012) Effectiveness of genomic prediction of maize hybrid performance in different breeding populations and environments. *G3* 2:1427–1436
- Winkler CR, Jensen NM, Cooper M et al (2003) On the determination of recombination rates in intermated recombinant inbred populations. *Genetics* 164:741–745

- Wright SI, Bi IV, Schroeder SG et al (2005) The effects of artificial selection on the maize genome. *Science* 308:1310–1314
- Yan J, Shah T, Warburton ML et al (2009) Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLoS ONE* 4:e8451
- Yan JB, Kandianis CB, Harjes CE et al (2010) Rare genetic variation at *Zea mays crtRB1* increases beta-carotene in maize grain. *Nat Genet* 42:322–327
- Yang XH, Yan JB, Shah T et al (2010) Genetic analysis and characterization of a new maize association mapping panel for quantitative trait loci dissection. *Theor Appl Genet* 121:417–431
- Young ND (1999) A cautiously optimistic vision for marker-assisted breeding. *Molec Breed* 5:505–510
- Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* 17:155–160
- Yu J, Holland JB, McMullen M, Buckler ES (2008) Genetic design and statistical power of Nested Association Mapping in maize. *Genetics* 178:539–551
- Yu JM, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zamir D (2001) Improving plant breeding with exotic genetic libraries. *Nat Rev Genet* 2:983–989
- Zhang M, Montooth KL, Wells MT et al (2005) Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* 169:2305–2318
- Zhang NY, Gibon Y, Gur A et al (2010a) Fine quantitative trait loci mapping of carbon and nitrogen metabolism enzyme activities and seedling biomass in the maize IBM mapping population. *Plant Physiol* 154:1753–1765
- Zhang ZW, Ersoz E, Lai CQ et al (2010b) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42:355–362
- Zheng P, Allen WB, Roesler K et al (2008) A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat Genet* 40:367–372

Chapter 26

Breeding Forest Trees by Genomic Selection: Current Progress and the Way Forward

Dario Grattapaglia

Contents

26.1	Introduction	652
26.2	Genomic Selection: Basic Concepts	655
26.2.1	Prediction Accuracy of Genomic Selection	658
26.2.2	Extent of LD as Main Determinant of Prediction Accuracy	658
26.2.3	LD and Marker Genotyping Density	659
26.2.4	LD and Effective Population Size of the Breeding Population	661
26.2.5	Genetic Architecture of the Target Trait	662
26.2.6	Training Population Size	663
26.3	Genomic Selection: Experimental Results In Forest Trees	664
26.4	Perspectives of GS in Forest Tree Breeding	667
26.5	Challenges of GS in Forest Tree Breeding	669
26.5.1	What Analytical Approaches to use for Genome-Based Prediction?	670
26.5.2	Will GS Be Able to Predict Non-Additive Genetic Effects?	671
26.5.3	What Will be the Accuracy of Predictions as Generations of GS Advance?	672
26.5.4	Will GS Models Work Across Different Populations?	674
26.5.5	Will GS Models Work Across Different Environments?	675
26.6	Conclusions	676
	References	678

Abstract A challenge common to all forest tree improvement programs is the long time interval of a breeding cycle. Moreover, the large size of trees, late trait expression and the extended time-lag between the breeding investment and the deployment of genetically improved material, make tree breeding a costly operation, more susceptible to changes in market demands, business objectives and climate change. The outlook of accelerating tree breeding and improving selection precision by marker

D. Grattapaglia (✉)
EMBRAPA Genetic Resources and Biotechnology—EPqB,
Brasilia, DF, 70770-910, Brazil
e-mail: dario.grattapaglia@embrapa.br

Universidade Catolica de Brasilia- SGAN,
916 modulo B, Brasilia, DF, 70790-160, Brazil

assisted selection (MAS), thus became one of the driving principles of most forest tree genome projects. Although important advances were made in quantitative trait locus (QTL) mapping and association genetics, MAS did not make it in the ‘real tree breeding world’. Limitations of early genomic technologies, coupled to the genetic heterogeneity of tree species and an overoptimistic assessment of the architecture of complex traits in such phenotypically plastic perennial organisms, largely explain this outcome. The inability to ascertain and make use of individual QTLs has caused a paradigm shift from trying to dissect trait components and determine their individual effects, to dealing with the aggregate of whole-genome effects to predict phenotypes by Genomic Selection (GS). Given the rapidly growing interest of tree breeders on this theme, this chapter provides an update on the current status and upcoming perspectives of GS in forest tree breeding. After a brief explanation of the basic principles and the main factors that impact prediction accuracy, the perspectives and the encouraging experimental results of GS in forest trees are reviewed. Concerns raised by tree breeders about GS are then discussed by reviewing the current knowledge in other species, while attempting to provide a roadmap for upcoming research and operational applications of GS. The prospects of GS in tree breeding are very promising to increase genetic gain per unit time through improved estimation of breeding (parent selection) and genotypic (clone selection) values, reduction of generation time and optimization of genome-directed mate allocation. Furthermore, the progressive accumulation of huge genotype and corresponding phenotype datasets in GS will provide an exceptional ‘big data’ framework that should enhance our understanding of the connection between genome-wide elements and the observable phenotypic variation in complex traits.

Keywords Genome-wide Selection · GWS · Eucalyptus · Whole-genome prediction · Tree breeding

26.1 Introduction

Tree breeding has become a key element of intensive forest-based operations worldwide, supplying genetically improved seeds and clones to increase the economic value of planted forests. Advanced tree breeding involves a large number of activities around the basic concept of recurrent selection aimed at increasing the frequency of favorable alleles for a number of traits simultaneously in the population. Repeated cycles of selection, inter-mating and genetic testing are used to develop genetically superior planting material in an economically efficient way by maximizing genetic gain per unit time at the lowest possible cost. Depending on the biology of the species, deployment plan, whether seeds or clones, and long term economic objectives of the forest products, significant differences will exist among programs both in breeding strategy, tactics and intensity (Namkoong et al. 1988; White et al. 2007).

All tree breeding programs, however, face a common challenge: the long time interval of a typical breeding cycle which may last several years to decades. Additionally, the large size of trees and the late expression of most important traits, make tree breeding a costly operation. The extended time-lag between the breeding investment and the ultimate deployment of genetically improved material also makes this endeavor susceptible to changes in market demands, business objectives and management policies. The uncertainty associated with planning and conducting decade-long breeding programs can be high. Not surprisingly, a significant effort has been devoted historically by tree geneticists and breeders to understand juvenile-mature correlations for complex traits such as height growth, wood properties and disease resistance (Namkoong et al. 1988). From those correlations ways have been devised to accelerate tree breeding by selecting trees as young as possible (Williams 1988) and by employing breeding procedures such as flower induction (Greenwood et al. 1991; Hasan and Reid 1995) to shorten the time required to recombine selected individuals.

It was in the perspective of accelerating tree breeding by early selection that molecular marker technologies raised alertness among tree geneticists early on. Besides shortening breeding cycles, early marker assisted selection (MAS) was perceived as a means to increase selection intensity, reduce effort of field-testing and improve selection precision for low heritability traits such as volume growth and late expressing ones such as wood properties. The long breeding cycles, the costs involved in establishing, maintaining and phenotyping large progeny trials for several traits of low heritability, were identified as the obstacles that MAS could help overcome (Neale and Williams 1991; Grattapaglia et al. 1992; Williams and Neale 1992). Nonetheless, since the fundamental principle of MAS is the existence of linkage disequilibrium between marker and QTL alleles, the prospects of MAS were challenged from its onset, due to the undomesticated nature and heterozygous state of forest tree populations in linkage equilibrium, together with concerns regarding QTL by background and by environment interactions, changes of QTL allele frequencies among generations and the genotyping costs involved (Strauss et al. 1992). Still, despite the validity of most of those issues, the appealing features of the application of MAS in tree breeding became the main rationale and driving force behind most, if not all, forest tree genome projects for the subsequent 20 years.

Pines, poplars and eucalypts have been the main targets of genome projects geared toward developing MAS tools since the early 1990s. Many promises have been made about the economic benefits derived from the incorporation of genome technologies into tree breeding. We were assured that we would be able to look at the tree's alleles at QTLs or genes and determine its breeding or genotypic value directly. This included the implicit, and maybe naïve, assumption that by the genetic dissection approach we would have already established the effects of all the important alleles during the life of the tree, in every population and environment. However, regardless of numerous advances in QTL and association mapping in forest trees, it seems that ascertaining these effects is proving much more elusive than originally assumed. In the meanwhile tree breeders have relentlessly kept their stride of mating 'the best to the best' and

making significant progress, boosted by the use of sophisticated quantitative genetics tools such as BLUP (Best Linear Unbiased Prediction) (Silva et al. 2000), high-throughput wood phenotyping methods such as NIRS (Near Infrared Spectroscopy) (Raymond and Schimleck 2002), improved flower induction procedures (Meilan 1997), hybrid breeding and refined nursery technologies for mass clonal propagation (Assis 2011). Despite shortcomings in our understanding of quantitative genetic variation, these methods are highly effective, although genetic evaluations use no explicit knowledge of individual gene action. A question may now be raised: 'Is such knowledge necessary to fulfill the expectations of MAS in tree breeding?'

Undoubtedly, considerable progress has been made in mapping QTLs and gene-trait associations in the main plantation forest tree species. Recent reviews have highlighted these developments (Grattapaglia and Kirst 2008; Grattapaglia et al. 2009; Neale and Kremer 2011; Harfouche et al. 2012). From the initial studies of single bi-parental populations of relatively limited size where apparently major, but mostly overestimated, effect QTLs were detected, recent studies with larger or multiple families have shown a different picture. As more individuals per family and more families are analyzed, the detection power increases, more QTLs are uncovered, the estimated effect of each one gets smaller and the inconsistency of these effects across backgrounds and environments becomes more evident (Dillen et al. 2008; Rae et al. 2008; Novaes et al. 2009; Thumma et al. 2010; Gion et al. 2011). With a larger number of QTLs controlling each trait, and each QTL with a small and unpredictably variable effect, the likelihood of implementing MAS for several traits simultaneously is practically precluded.

The inability to ascertain with any confidence the effects of QTLs has now caused a paradigm shift in molecular breeding, starting in domestic animals and rapidly permeating into crop and forest trees. We are now moving from trying to discover genes and determine their individual effects, back to dealing with the aggregate of the genes. Predicting the Genomic Estimated Breeding Value (GEBV) of an individual is what Genomic Selection (GS) or Genome-Wide Selection (GWS) is now pledging. Differently from the underlying principle of MAS, where a subset of well verified marker-trait associations are first discovered and then used for selection, GS estimates all marker effects simultaneously, retaining all of them as predictors of performance, thus precluding the prior search for significant marker-trait associations but focusing exclusively on selection efficiency (Meuwissen et al. 2001; Goddard and Hayes 2009).

In spite of the vast literature of simulation-based and experimental studies in domestic animals (Hayes et al. 2009b; Luan et al. 2009; VanRaden et al. 2009; Daetwyler et al. 2010) and a rapidly increasing treatment of this topic in crop (Janink et al. 2010; Lorenz et al. 2011; Nakaya and Isobe 2012) and tree genetics (Grattapaglia and Resende 2011; Iwata et al. 2011; Denis and Bouvet 2013; Kumar et al. 2012a), GS has not yet become popular in applied breeding mainly due to a general lack of digestible information to breeders and the still small number of empirical studies to date. Two studies have pioneered the experimental demonstration of GS in forest tree breeding, one in loblolly pine (*Pinus taeda*) (Resende et al. 2012b)

and one in eucalypts (mainly *Eucalyptus grandis*) (Resende et al. 2012a). Following those two reports in trees, two additional studies have been published in loblolly pine (Resende et al. 2012c; Zapata-Valenzuela et al. 2012) and one in the apple fruit tree (*Malus domestica*) (Kumar et al. 2012b).

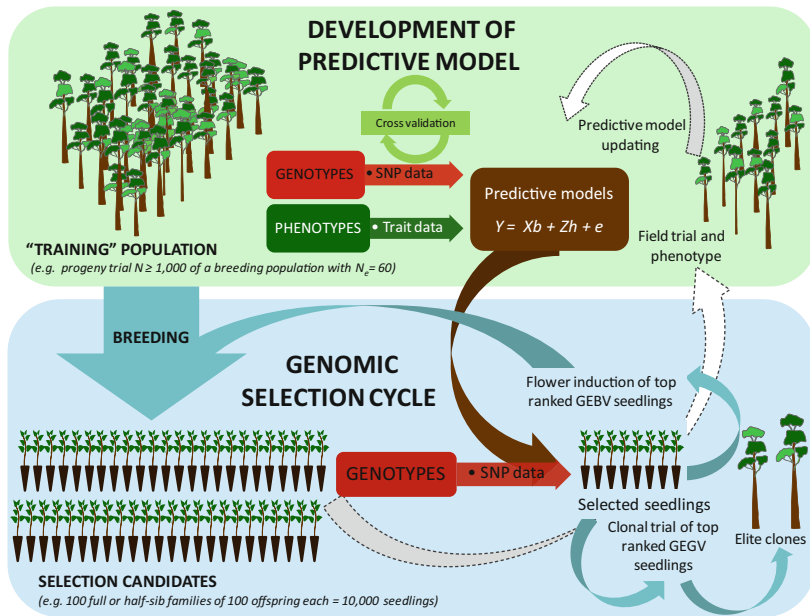
Given the rapidly growing interest of tree breeders on the theme, the objective of this chapter is to provide an update on the current status and upcoming perspectives of Genomic Selection in forest tree breeding, building upon the existing literature in animal and crop breeding. Initially a brief explanation is provided to understand the basic principles of GS and the main factors that impact its prediction accuracy as evaluated by recent simulation-based studies. No detailed treatment is given to the several statistical aspects of GS, but recent reviews with a thorough discussion of this topic can be found (Lorenz et al. 2011; de Koning and McIntyre 2012). Secondly the perspectives of GS in forest tree breeding and the encouraging experimental results of GS in trees are reviewed. Finally the main concerns and challenges generally raised by tree breeders about GS are discussed in detail by reviewing the current knowledge while attempting to provide a roadmap for upcoming research and initial applications of GS in forest tree improvement.

26.2 Genomic Selection: Basic Concepts

Genomic Selection (GS) or Genome-Wide Selection (GWS) was proposed 11 years ago as a new paradigm based on genome-wide marker assisted selection (MAS) that could substantially increase the rate of genetic gain in animals and plants, especially if combined with reproductive techniques to shorten the generation interval (Meuwissen et al. 2001; Goddard and Hayes 2009). This concept was based on technical advances and declining costs of high-throughput genotyping and new statistical methods for large datasets. The impact of higher density genotyping on marker assisted selection had been anticipated a few years earlier by Haley and Visscher (1998) when they said: *‘Emerging technologies could allow large numbers of polymorphic sites to be identified, practically guaranteeing that markers will be available that are in complete association with any trait locus. Identifying which polymorphism out of many that is associated with any trait will remain problematic, but multiple-locus disequilibrium measures may allow performance to be associated with unique marker haplotypes. This type of approach, combined with cheap and high density markers, could allow a move from selection based on a combination of “infinitesimal” effects plus individual loci to effective total genomic selection’*. GS is now a reality in routine animal breeding (Hayes et al. 2009b; Hayes and Goddard 2010; Pryce and Daetwyler 2012) and became a topic of great interest in plant breeding in the last 5 years starting with the influential papers by Bernardo and Yu (2007) and Bernardo (2008) followed by others (Grattapaglia et al. 2009; Heffner et al. 2009; Jannink et al. 2009).

Although both MAS and GS start with a ‘discovery’ phase, where relationships between genotypes and phenotypes are established in relevant populations using specific statistical approaches, GS is fundamentally different from MAS. MAS focuses on a limited number of marker-trait associations that are first discovered in one or a few families or association mapping panels using rigorous significance tests, and then used for selection. GS on the other hand uses a dense genome-wide panel of markers whose effects on the phenotype are estimated simultaneously in a large and representative population of individuals without applying rigorous significance tests, but rather retaining all or a large proportion of markers as forecasters of phenotype in prediction models. In GS a marker effect does not need to exceed a stringent significance threshold to be used in the subsequent breeding phase and the effects of the marker alleles are estimated in a larger population rather than within one or a few mapping families. GS therefore works on the principle that linkage disequilibrium (LD), provided by dense genotyping, is sufficient to track all relevant QTL effects for the target trait which are expected to be in LD with at least some of the queried markers. By avoiding prior marker selection and estimating marker effects in a large and representative ‘training’ population, GS tends to capture most genetic variance for the trait mitigating the quandary of how to capture the “missing heritability” of complex traits likely explained by large numbers of small effects that QTL or association genetics-based MAS does not capture.

In GS a ‘training’ or ‘estimation’ population, used to estimate marker effects involves several hundreds to a few thousand individuals representative of the reference breeding population which are genotyped for a genome-wide panel of markers and phenotyped for all traits of interest. From these data sets, prediction models are developed for each trait and cross validated by means of a ‘validation’ population, a randomly sampled subset of individuals of the same training population that did not participate in the estimation of marker effects. Once a prediction model is shown to provide satisfactory selection accuracy, i.e. correlation between the observed and predicted breeding values obtained by cross-validation, it can be used in the breeding phase to calculate the genomic estimated breeding values (GEBV) or total genomic estimated genotypic values (GEGV) (when non-additive effects are also targeted) of the selection candidates, i.e. a set of individuals for which only genotypes are recorded and phenotypes are to be predicted by the breeder (Box 1). A GEBV is calculated by multiplying the number of alleles at all markers by their effect estimated by, for example, random regression best linear unbiased prediction (RR-BLUP) or any other statistical method that adequately avoids model over-fitting by marker-specific shrinkage of regression coefficients (Crosa et al. 2010; Lorenz et al. 2011). GS consequently produces a single breeding or genotypic value for each individual, moving away from the classical trait dissection approach by de-emphasizing the contribution of individual genes or QTLs to the target quantitative trait.



Box 1. General Scheme of the Steps Involved in Developing Genomic Selection in a Tree Breeding Program

The top chart depicts the development of the predictive model from a training population that was genotyped and phenotyped and used for cross validation. This can typically be a progeny trial ($N \geq 1,000$) derived from inter-mating elite parents that form a breeding population with effective population size (N_e) between 30 and 100. The bottom chart illustrates the Genomic Selection cycle starting with breeding elite parents to generate the selection candidates, typically an array of full or half-sib families. These are genotyped and have their breeding values (GEBV) and genotypic values (GEGV) estimated using the predictive model developed earlier. Top ranked seedlings for GEBV are subject to early flower induction and inter-mated to create the next generation of breeding. Top ranked seedlings for GEGV are clonally propagated and tested in verification clonal trials where elite clones are eventually selected for operational plantation. Additionally, as the white dotted arrows indicate, a random subset of the selection candidates are planted in experimental design and phenotyped after a few years to provide genotype and trait data for GS model updating as generations of GS are carried out, mitigating the erosion of marker-QTL LD, therefore maintaining accuracy of GS predictions over generations.

26.2.1 Prediction Accuracy of Genomic Selection

The prediction accuracy of a GS model is evaluated by the correlation between the GEBV and the experimentally estimated breeding values (EBV), where the EBV can be obtained in a number of ways, most simply, as a phenotypic mean. The prediction accuracy of GS basically depends on four parameters (Hayes et al. 2009b): (1) the extent of LD between markers and QTLs; (2) the number of individuals with phenotypes and genotypes in the training population from which the marker effects are estimated; (3) the heritability of the trait in question; and (4) the distribution of QTL effects (number of loci and size effects). The first two parameters can be controlled experimentally by the breeder, while the other two are intrinsic to the genetic architecture of the trait of interest in the specific environment and genetic background of the target population and have to be lived with by the breeder when devising the best GS strategy.

These four main parameters known to impact the accuracies of GS are evidently strongly interconnected and interdependent. Still, an attempt to assess the impact of each one individually was recently reported under a set of realistic circumstances in terms of genotyping technologies and effective population sizes used in tree breeding. Based on deterministic simulations, some broadly useful guidelines emerged to stimulate the discussion of GS for tree breeding, regardless of the target species, recombinant genome size, or breeding cycle length (Grattapaglia and Resende 2011). A similar discussion was presented in the context of apple breeding (Kumar et al. 2012a). Overall the results of those two reports matched those obtained in previous simulations studies reviewed for different animal and crop breeding scenarios (Godard and Hayes 2009; Lorenz et al. 2011), although specific issues relevant to forest tree breeding should be pointed out as described below.

26.2.2 Extent of LD as Main Determinant of Prediction Accuracy

The main issue that will largely determine the prediction accuracy of GS is the extent of linkage disequilibrium, i.e. the non-random association between marker alleles and QTL alleles. This quantity in turn depends on the effective population size (N_e) and the number of markers used. The effective population size corresponds to the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration (Wright 1931). As the effective population size gets smaller, the effect of genetic drift gets stronger and more LD is generated because it is unlikely that combinations of marker alleles and QTL alleles get sampled at a frequency that corresponds to the product of their individual frequencies. The resulting non-random association between alleles at marker loci and QTLs allow marker alleles to predict the allelic state of nearby QTL, and thus to predict phenotypes. At equilibrium, the LD generated by random drift is balanced by recombination that takes place as breeding generations advance, causing it to dissipate, such that closer loci are expected to be in higher LD than more distant ones. As a

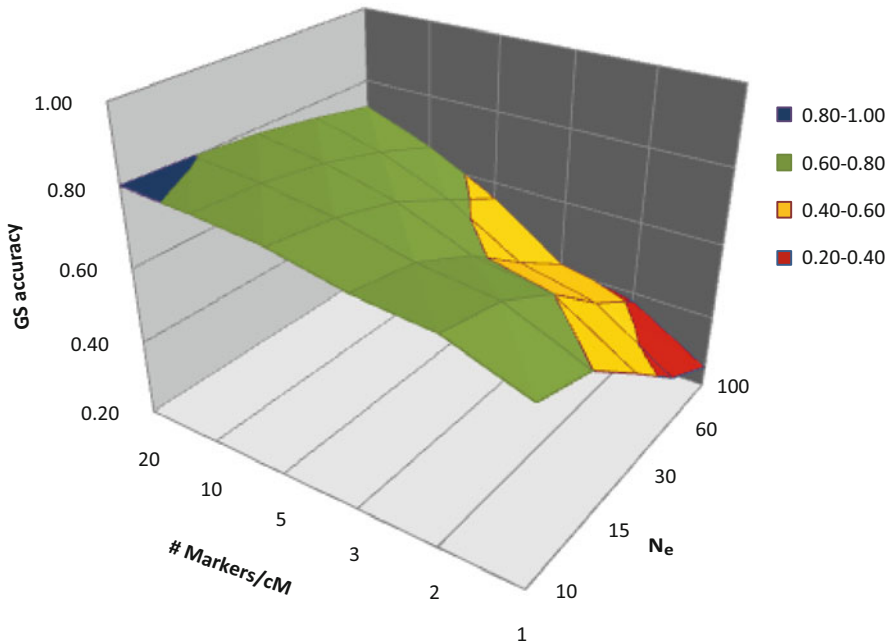


Fig. 26.1 The simultaneous impact of effective population size (N_e) and genotyping density (markers/cM) on the accuracy of Genomic Selection assuming a trait heritability of 0.2, a training population of $N = 1000$ individuals, and 100 QTLs controlling trait variation. The green and blue surfaces denote satisfactory to excellent ranges of accuracy, respectively.

consequence, the relationship between N_e and LD impacts the marker density needed to achieve and sustain adequate prediction accuracy of a GS model across generations. In other words, marker density needs to scale with the effective population size and the level of LD between markers and QTL can be increased by reducing N_e . The good news is that both N_e and marker density can be controlled by the breeder.

26.2.3 LD and Marker Genotyping Density

The extent of marker–QTL LD, modeled by varying the effective population size and genotyping density under the Sved equation (Sved 1971), showed by far the largest impact on the prospects of GS in forest tree breeding (Grattapaglia and Resende 2011). The simultaneous impact of these two parameters on the accuracy of GS can be visualized in a surface 3D graph. Assuming a trait heritability of 0.2, a training population of $N = 1000$ individuals, and 100 QTLs controlling trait variation, the green surface denotes a satisfactory range of accuracy > 0.60 , while the blue one indicates accuracy way above what typically reached by conventional phenotypic BLUP selection (Fig. 26.1). The upper bound benchmark accuracy of phenotypic BLUP selection, set at 0.68 in that study, can be reached at a relatively low marker

density, around 2–3 markers/cM, as long as the effective population size is kept below $N_e = 60$. For larger effective population sizes up to $N_e = 100$, however, 10 or up to 20 markers/cM would be necessary for keeping high accuracies of GS. Such a target genotyping density will require the development of genotyping arrays to yield somewhere between 20,000 and 50,000 informative markers depending on the size of the recombining genome and the final effective population size of the breeding population.

In *Eucalyptus* with a recombining genome extending between 1,100 cM in a multi-species high density reference map (Hudson et al. 2012) and 1,500 cM for *E. grandis* (Brondani et al. 2006), 22,000–30,000 informative markers would suffice, while in loblolly pine (*Pinus taeda*) with estimates varying between 1,500 cM (Echt et al. 2011) and 1,900 cM (Sewell et al. 1999), 30,000–40,000 markers would be necessary. Currently, reaching such numbers of markers does not represent a problem. Discovery of several hundreds of thousands of SNPs can be readily achieved using deep (> 20–30X coverage) resequencing of complete genomes or reduced representation libraries of germplasm-representative samples of individuals, as shown in plants and animal species (Baird et al. 2008; Van Tassel et al. 2008; Myles et al. 2010). This strategy proved effective in a preliminary study in *Eucalyptus* where > 200,000 high quality SNPs were discovered by RAD (Restriction Associated DNA) sequencing, out of which 42,000 were simultaneously polymorphic in the two most widely planted and phylogenetically divergent *Eucalyptus* species (Grattapaglia et al. 2011). Recently, the boost in throughput achieved by next-generation DNA sequencing, allows SNP genotyping directly on reduced genomic representations of barcoded individual samples sequenced in multiplexed pools. Such genotyping-by-sequencing (GbS) approaches have been applied to inbred crops (Elshire et al. 2011; Poland et al. 2012) and show promise in outbred *Eucalyptus*, although with additional challenges to reach adequate repeatability and call rates for routine large scale use (Sansaloni et al. 2011; Faria et al. 2012).

The impact of the genotyping density used in the practice of GS will become even more important as generations of selection advance. In the absence of selection, increasing marker density is beneficial to the persistence of GEBV prediction accuracy over generations (Solberg et al. 2009), because higher marker densities enable GEBV accuracy to persist over time due to a slower decay of LD among tightly linked marker and trait loci. However, directional selection following the initial training population is expected to result in a rapid decline of accuracy (Muir 2007). High-density genotyping was shown to be essential to sustain accuracy and keep selection effective for more generations in the presence of directional selection when a finite number of QTL loci is assumed rather than an infinitesimal model (Long et al. 2011). In such cases, selection, together with recombination, may change the pattern of LD between markers and QTLs. The new LD generated by selection can be unfavorable for GEBV prediction which was based on the original marker–QTL LD structure in the training population.

Although the decrease of accuracy of GS over time can be mitigated by re-estimating marker effects or varying the weight given to markers (see below), the possibility of using higher genotyping densities, provided that costs are kept

affordable, should always be a priority. In view of the genetic heterogeneity and undomesticated nature of forest trees genomes, attempts to use reduced marker panels as an option to reduce genotyping cost as proposed for domestic animal and crops (Habier et al. 2009) should be seen with caution. Notwithstanding the fact that the structure of LD in trees is different than animals and crops, a lower marker density would make GS considerably more susceptible to the decay of LD with recombination and selection. Additionally, multi-trait selection or the adoption of GS in variable populations will result in different sets of markers being fitted in predictive models. In such a scenario a high density marker panel useful across variable breeding populations and aimed at selecting for several traits simultaneously will by far be the best option. Current genotyping technologies, both high-density SNP chips and genotyping-by-sequencing protocols, are moving toward a situation where the cost increase for a genotyping density increase of one or two orders of magnitude more markers will become less of an issue. The ‘per sample’ cost, provided that an adequate number of markers is obtained for the envisaged population and traits, will be the key issue.

26.2.4 LD and Effective Population Size of the Breeding Population

Once an adequate genotyping density can be achieved at a reasonable cost, the opportunity to carefully control the effective population size of a breeding population is a distinct advantage in forest trees when compared to most animal breeding settings where GS is currently used. The random genetic drift that operates when a breeding population is established by selecting elite parents from a natural population generates new LD. In other words, marker-QTL associations in linkage equilibrium in the ancestral natural population go into LD in the smaller breeding population. Calibrating the extent of LD by managing the effective population size, such that near-maximum genetic gain can be achieved in a long-term breeding program, is thus a key element when adopting GS.

Based upon theoretical studies and practical considerations regarding the appropriate size of a tree breeding population, populations with N_e between 20 and 50 have been recommended to support selection and breeding programs with appreciable genetic gains for several generations (Namkoong et al. 1988; White et al. 2007). Although suitable for short-term genetic gains, such constrained N_e are however subject to larger deviations of actual versus predicted progress and may result in a faster build-up of relatedness. To sustain genetic gains in long term tree breeding programs, effective population sizes between $N_e = 40\text{--}100$ are generally used, typically corresponding to a census number (i.e. the total number of selections retained in the breeding population in any given generation) around 200–300 individuals with some level of relatedness (White et al. 2007). As examples, the third breeding cycle of Loblolly Pine in the Southeastern U.S., has adopted a highly selected group of 40 selections to provide rapid gains (McKeand and Bridgwater 1998). In *Eucalyptus*,

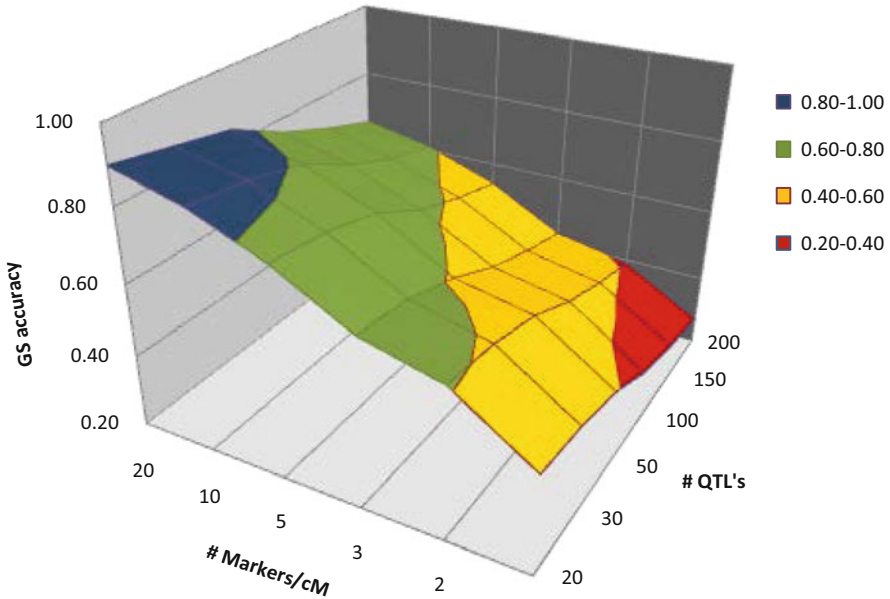


Fig. 26.2 The simultaneous impact of the numbers of QTLs controlling trait variation and genotyping density (markers/cM) on the accuracy of Genomic Selection assuming a trait heritability of 0.2, a training population of $N = 1000$ individuals, and an effective population size $N_e = 60$. The green and blue surfaces denote satisfactory to excellent ranges of accuracy, respectively.

populations with N_e between 10 and 60 are typically used in the advanced strategy called Reciprocal Recurrent Selection between Synthetic Populations, an approach that exploits the variation derived from multiple species for hybrid breeding (Assis and de Resende 2011). Thus, the effective population sizes currently used in most tree breeding programs largely fit within the perspectives of reaching high prediction accuracy by GS provided that adequate genotyping densities are used (Fig. 26.1).

26.2.5 Genetic Architecture of the Target Trait

Besides the extent of marker–QTL LD, the number of QTLs controlling trait variation has a distinct impact on the accuracy of GS. Fewer loci controlling larger fractions of the phenotypic variance are more easily captured relative to a more complex genetic architecture involving larger numbers of loci. As expected, the reduction of GS accuracy with an increasing number of QTLs involved tends to be more pronounced at lower marker densities or larger effective population sizes. For a fixed effective population size of $N_e = 60$, trait heritability of 0.2 and a training population of 1,000 individuals, the simultaneous impact of genotyping density and the number of QTLs on the accuracy of GS can be visualized in a 3D surface graph (Fig. 26.2) built from previous simulation results (Grattapaglia and Resende 2011). Given the QTL mapping results that emerged from better powered experiments (i.e. larger family sizes and multiple families), several tens of QTLs likely control each single trait in

forest trees (Dillen et al. 2008; Gion et al. 2011; Novaes et al. 2009; Rae et al. 2008; Thumma et al. 2010). It is reasonable, therefore, to assume that a quantitative trait will be controlled by at least 20 and up to 100 or more QTLs, while 200 QTLs would be a minimum number when several independent traits would be tackled concurrently by GS. Under such circumstances, satisfactory to excellent accuracies of GS (green and blue surfaces) would be reached with marker densities ≥ 5 markers/cM assuming a simpler genetic architecture, while 20 markers/cM would be necessary with larger numbers of QTLs.

Heritability, as an additional factor underlying the genetic architecture of a trait, was shown to have a relatively minor impact on accuracy under the assumptions used to simulate the expected performance of GS in tree breeding (Grattapaglia and Resende 2011). Because GS accuracy was calculated by a deterministic approach, it is directly proportional to the product of the heritability and the ratio between the number of phenotypic records in the training population and the number of QTLs involved as described earlier (Daetwyler et al. 2008). Therefore by using a rather modest training set for a tree breeding situation of $N \geq 1,000$ individuals, a trait controlled by 100 QTL, and an effective population size $N_e = 60$, the GS accuracy increased only slightly, from 0.71 to 0.83, as the heritability went from 0.2 to 0.6 (Grattapaglia and Resende 2011). Simulation studies for animal breeding scenarios also showed that a decrease in accuracy with decreasing heritability can be readily compensated by using larger training sets (Meuwissen et al. 2001; Nielsen et al. 2009). A similar small impact of increased heritability on the accuracy of GS was also seen in simulations in the context of apple breeding (Kumar et al. 2012a).

26.2.6 Training Population Size

While it can be a challenging endeavor in animal breeding (Goddard and Hayes 2009), assembling a large number individuals into a training population to accurately estimate markers effects is generally not a limitation for forest trees. Choice of a training population evidently will depend in large part on the breeding strategy adopted and the number and structure of populations involved. Training populations generally can be established by sampling trees in existing progeny trials derived from the inter-mating (open pollinated or controlled) of a set of a few dozen elite parents representative of the target genetic variation with adequate effective population size to provide sustained gains for a few generations ahead. Combining training sets from different populations can be useful to boost accuracy when individual populations lack sufficient size although considerable risks exist of lowering the performance of such multi-population prediction models when genetic backgrounds are different. Furthermore, with current drops in genotyping costs, while phenotyping costs remain constant or increase, efforts might be reversed and individuals to be phenotyped for a training set can be chosen on the basis of their genotype. Nevertheless, costs of phenotyping some wood properties traits in training sets of hundreds or thousands of trees warrant optimization using, for example, indirect methods such as NIRS (Near Infrared Reflectance Spectroscopy) as recently done in an experimental study in *Eucalyptus* (Resende et al. 2012a).

With up to $N = 1,000$ individuals used to train a GS model, the selection accuracy was shown to rapidly increase, reaching satisfactory levels. Using 2,000 individuals as training set, a small improvement of 6–10 % of the accuracy achieved with $N = 1,000$ was seen. After $N = 2,000$ the accuracy tends to plateau irrespective of the effective population size and genotyping density (Grattapaglia and Resende 2011). However, if the QTL distribution violates the infinitesimal model assumption of equal size effect and common variance, not all of the genetic variance is explained and the selection accuracy can be lower depending on the method used to calculate the GEBV (Coster et al. 2010). Using training sets larger than $N = 2,000$ might, therefore, be warranted to protect against such model violations or cases where several hundred QTLs control trait variation. Furthermore, larger training populations mitigate the probability of losing rare favorable alleles from the breeding population as generations of selection advance, although some will inevitably be lost because they are in low LD with any marker. A higher marker density will also help in this respect, i.e. in preserving rarer alleles in the breeding populations, thus allowing better long term gains from selection.

26.3 Genomic Selection: Experimental Results In Forest Trees

The first experimental results of GS for animal and plant species in general were reported around the same time, between 2007 and 2008. A recent review compiled detailed tables with these results until mid 2011 (Nakaya and Isobe 2012). Although the demonstration of GS was pioneered in bovines and reported in conference proceedings as described by Hayes et al. (2009b) and Pryce and Daetwyler (2012), the first formally published study was in mice. The use of 10,946 SNP markers in a heterogeneous mouse population of 1,884 individuals increased the accuracy of prediction of genetic values up to 0.57 (across family) and 0.14 (within family) (Legarra et al. 2008). In plants the first assessments of GS were done in relatively small biparental populations of maize, barley and Arabidopsis. Accuracies between 0.31 and 0.83, depending on the species and trait, were reported (Lorenzana and Bernardo 2009). After those initial studies, several other followed in cattle and chicken reporting accuracies usually between 0.1 and 0.7 depending on training population size, number of markers and trait (Nakaya and Isobe 2012). In crop plants seven other studies followed that initial account, three of them (Albrecht et al. 2011; Riedelsheimer et al. 2012; Zhao et al. 2012) not yet listed in the review tables of (Nakaya and Isobe 2012). All these studies were carried out either in maize or wheat reporting considerably higher accuracies in the range of 0.4 and 0.9.

Reports of experimentally estimated accuracies of GS in forest tree species promptly followed those from crop plants. The fact that BLUP methodologies and the concept of breeding values, both derived from animal breeding, are well established in forest trees, likely contributed to a quick recognition of GS as a potential breeding tool in forest trees. Experiments in forest trees have been very distinctive in the fact that considerably larger training population sizes (several hundred individuals instead of a few hundred) and numbers of markers (a few thousand instead of a few hundred) were used when compared to crop plants experiments. Reported

accuracies in forest trees, however, have been in the same range as those in crop plants. This fact substantiates the narrow genetic diversity of crops as a result of a small number of founder lines coupled to self reproduction, conditions that generate extensive LD. On the other hand forest trees are essentially undomesticated, genetically very diverse with shorter range LD, requiring larger training populations and higher marker densities as predicted by theory.

Results of GS studies in forest trees are summarized (Table 26.1). A study in *Eucalyptus* involving two independent breeding populations (Resende et al. 2012a) and a second one in a loblolly pine population (Resende et al. 2012b) led the way in GS of forest trees. Accuracies averaging 0.5–0.8 were estimated by cross validation for all traits, with a few exceptions. These results approximated quite well to the accuracies predicted from deterministic (Grattapaglia and Resende 2011) and stochastic simulations (Iwata et al. 2011) for similar parameters of trait heritability, effective population size and genotyping density. Potential gains of 50–200 % in selection efficiency predicted by simulations could therefore be corroborated by these experimental reports. Recently two additional experimental reports in loblolly pine were added to this list, one as a follow up of the same population genotyped earlier (Resende et al. 2012b) but analyzing several additional traits under different statistical models (Resende et al. 2012c), and a second one in a small and highly structured population of loblolly pine (Zapata-Valenzuela et al. 2012). Although not a forest species, the first experimental GS study in a fruit tree recently reported accuracies varying from 0.70–0.90 for various fruit quality traits using a population of 1,120 apple seedlings generated from a factorial mating design of four females and two male parents (Kumar et al. 2012b).

The proof-of-concept study in *Eucalyptus* (Resende et al. 2012a) demonstrated that GS not only achieved accuracies as good as, or better than, those attainable by conventional phenotypic selection for growth and wood quality traits, but also captured large fractions (75–97 %) of trait heritability that association genetics and QTL mapping classically fail to explain. Additionally that study asked a significant question to practical tree breeding: can a GS model fitted to one population be suitable to predict phenotypes in an unrelated population? The populations had a similar genetic background as far as species composition, although trees were genetically unrelated and grown in different environments. When the prediction models developed for one population were used to predict phenotypes in the second one, or vice versa, the GS accuracies declined drastically to values close to zero. Interestingly, however, while GS accuracies across populations were poor, a highly significant level of coincidence between the two populations was observed regarding the physical location of the genomic regions underlying the measured traits. These results indicated that in spite of a significant between-population conservation of the loci underlying the quantitative traits, the allelic effects vary across populations, making predictions inaccurate likely as a result of variable patterns of marker–QTL LD, inconsistent allelic effects in different backgrounds and genotype by environment interaction. These experimental results showed that GS prediction models will likely be population-specific, although multi-population GS models might be feasible with higher genotyping density so that marker–QTL linkage phase would persist across populations. However, the genotype by environment interaction might supplant the persistence of LD relationships and cause equally unacceptable accuracies.

Table 26.1 Summary of experimental results of genomic selection in forest trees

Species	Population type	Training population size	# and type of markers used	Trait	Accuracy of GEBV	Reference
<i>Eucalyptus grandis</i> <i>x E. urophylla hybrids</i>)	Progeny trial of 43 full-sib families from 11 elite hybrid parents	738	3,129 DArT	Circumference growth Height growth Wood specific gravity Pulp yield	0.74 0.79 0.78 0.88	(Resende et al. 2012a)
<i>Eucalyptus (E. grandis, E. urophylla, E. globulus and F1 hybrids of these species)</i>	Progeny trial of 232 full-sib families from 51 elite parents	920	3,564 DArT	Circumference growth Height growth Wood specific gravity Pulp yield	0.73 0.66 0.65 0.55	
Loblolly pine <i>Pinus taeda</i>	Progeny trial of 61 full-sib families from 32 parents	800	4,852 SNP	Diameter growth Height growth	0.65–0.73 ¹ 0.64–0.74	(Resende et al. 2012b)
Loblolly pine <i>Pinus taeda</i>	Progeny trial of 61 full-sib families from 32 elite parents	951	4,853 SNP	Growth (several traits) Development (several traits) Fusiform rust resistance Wood stiffness Lignin content Latewood %	0.38–0.49 ² 0.24–0.51 0.23–0.34 0.39–0.43 0.17 0.23–0.24	(Resende et al. 2012c)
Loblolly pine <i>Pinus taeda</i>	13 full-sib families related by a common parent	149	3,406 SNP	Wood specific gravity 5- and 6-carbon sugar content Lignin content Cellulose content Height growth Volume growth	0.20–0.22 0.25–0.26 0.66–0.76 0.61–0.83 0.47–0.52 0.30–0.56	(Zapata-Valenzuela et al. 2012)

Two key issues to forest tree breeding, GS accuracy across ages and environments, were assessed in the loblolly pine study (Resende et al. 2012b). Firstly, because diameter and height growth measurements were obtained over multiple years, separate prediction models could be built for each year. As expected, given the weak juvenile-mature correlations typically observed in conifers (Namkoong et al. 1988), GS models developed based on phenotypes measured early in the life of the trees (age 1 year and 2 year) had unacceptable accuracy in predicting phenotypes at age 6 year. Secondly, as the training population was clonally replicated across a wide north-south transect in southeastern USA, prediction models could be developed using the same genotype dataset but different phenotype datasets for the different environments. As expected, the GS accuracies of cross validation were high when applied within the same site, but declined considerably when used to predict growth in different sites. The decrease in accuracy paralleled the increase in geographic distance and latitude between the site for which models were estimated, and the site where they were validated. These results substantiate the fact that for accurate prediction of tree growth, phenotypes used in GS model development have to be measured in the training population at the same age and in the same or close breeding zone to the one where trees will be harvested.

26.4 Perspectives of GS in Forest Tree Breeding

Simulation-based and experimental reports outlined the promising prospects of GS to increase the efficiency of tree breeding programs. This would be accomplished fundamentally by shortening the length of the breeding cycle after precluding the progeny testing phase while practicing ultra-early selection for yet-to-be observed phenotypes at the seedling stage (Box 2). These selected juveniles could then be induced to flower either by top-grafting or by chemical treatment and recombined to ultimately conclude a cycle of genetic improvement several years earlier. In eucalypts and poplars GS not only could eliminate the progeny trial but would also reduce the time and costs involved in the clonal testing phase by reducing the number of selected trees that are tested as clones in a preliminary, typically large scale, clonal trial. In conifers, as pointed out earlier (Resende et al. 2012b), GS combined to somatic embryogenesis (SE) could considerably boost the efficiency of current clonal propagation protocols by allowing pre-selection of zygotic embryos based on their GEBV and their immediate expansion into elite SE lines for the establishment of clonal trials or directly into commercial plantations.

Besides the time gain, a less mentioned advantage of GS is related to the possibility of efficiently carrying out selection for several traits simultaneously in large numbers of individuals. It is virtually impossible to any breeding program to complete a rigorous assessment of wood volume, stem taper and straightness, wood properties, sprouting and rooting abilities, nutritional efficiency, tolerance to pests and diseases, drought and frost, for all trees in a progeny trial. Even in clonal trials this is typically accomplished only in the very final stages of clonal trials for a very limited number of clones (20–50) that had been pre-selected for volume growth.

Box 2. Comparative Timelines of Genomic Selection (GS) Breeding and Phenotypic Selection (PS) Breeding for Tropical Eucalyptus

Both methods start at year zero with the same breeding population, although in GS it is assumed that predictive models have been previously developed as described in Box 1. In a GS breeding cycle, following SNP genotyping and genomic estimations for all target traits (i.e. growth, form, wood properties, disease resistance, etc.), selection candidates can follow three possible non-exclusive routes: (1) top ranked seedlings for GEBV (Genomic Estimated Breeding Value) are immediately routed to flower induction treatment and recombined to create the improved population (green boxes) completing the recurrent breeding cycle; (2) top ranked seedlings for GEGV (Genomic Estimated Genotypic Value) are cloned directly by mini-cutting methods and directed to a verification clonal trial and ultimately submitted to a final selection for elite operational clones for plantation (red boxes); (3) a random subset of a few hundred or even all selection candidates in each GS cycle can be planted in the field in an experimental design to provide additional phenotypic data in due course which, together with the already collected genotypic data, will allow continuous predictive model updating (grey boxes). In summary GS precludes the progeny testing phase, accelerates the completion of a breeding cycles when coupled to early flower induction and allows reaching elite clones much faster. With GS, a cycle of recurrent selection, going from an original population to an improved population, lasts 5 years, while in conventional breeding it lasts 10 years. Two generations of elite clones can be reached by GS in 14 years while PS will only provide one generation in 15 years. Note that in PS the verification clonal trial lasts six to seven years to allow adequate phenotyping of wood properties traits. In GS accurate predictions of wood properties traits are obtained by GEBV so that a 5–6 year verification clonal trial serves mostly to validate the general field performance of the clones.

26.5 Challenges of GS in Forest Tree Breeding

GS research in forest trees is still in its infancy. As pointed out earlier, some issues still require careful examination and additional experimental data before GS can be adopted operationally (Grattapaglia and Resende 2011; Iwata et al. 2011). Some of the fundamental issues regarding the design of training populations, effective population size and genotyping density have been discussed above. Some other logistic issues such as specific nursery infrastructure, sample collection and tracking system, large scale DNA extraction and qualification, genotyping service providers and data analysis pipelines, are equally important for the successful implementation of a GS

operation but beyond the scope of this chapter. The discussion below focuses on the main standing questions about GS typically raised by breeders. These are briefly discussed taking into account what is currently known in an attempt to stimulate brainstorming and provide a roadmap for upcoming research and application of GS in forest tree improvement.

26.5.1 What Analytical Approaches to use for Genome-Based Prediction?

Whole-genome prediction requires methods that are capable of handling cases where the number of marker variables greatly exceeds the number of individuals, mitigating the risk of model over parameterization. Several analytical approaches have been proposed and used for prediction of genome estimated breeding or genotypic values. An optimal GS approach should provide the highest accuracy possible, limit overfitting on the training dataset, and be based as much as possible on marker–QTL LD rather than on pedigree relationships (kinship). Additionally, such methods must be easy to implement, reliable across a wide range of traits and datasets, and computationally efficient (Heslot et al. 2012). A discussion of GS methods is beyond the scope of this chapter. Several thorough reviews are available regarding the features of the main prediction methods for GS (Heffner et al. 2009; Lorenz et al. 2011) as well as comparative benchmark assessments in animal (Moser et al. 2009), crops (Heslot et al. 2012) and forest trees (Resende et al. 2012c).

The current approaches basically differ with respect to the assumptions regarding the genetic architecture of the trait for which genomic predictions are sought. For the scope of this chapter, a simplified view can be considered, grouping the approaches into three main categories: (1) shrinkage models such as ridge regression–best linear unbiased prediction (RR-BLUP), where the assumption made is that the trait is controlled by many loci of small effect, so that all marker effects are random, normally distributed and with a common variance. Instead of classifying markers as either significant or as having no effect, ridge regression shrinks all marker effects toward zero; (2) Bayesian methods such as BayesA and BayesB, where the presumably incorrect assumption of equal variance is relaxed and better modeling of marker effects of differing sizes is carried out. A separate variance is estimated for each marker, and the variances are assumed to follow a specified prior distribution. The assumption is that most markers have no effect on the trait and thus are left out of the prediction model; and (3) semi or non-parametric methods such as support vector regression and random forest regression that make no assumption of underlying genetic architecture, involve different fundamental theory when compared to linear models, and are expected to better capture non-additive effects in the prediction models.

Interestingly, experimental reports reviewed both in plants and animals have generally concluded that the RR-BLUP approach using a mixed model proves very effective in providing the best compromise between computation time and prediction

efficiency (Lorenz et al. 2011), suggesting that most quantitative traits adequately fit into the assumption of an infinitesimal model. In a loblolly pine study, for example, the performance of four different statistical methods was only marginally different when compared across 17 traits with distinct heritabilities and presumed genetic architectures, including growth, development and disease-resistance traits. A performance difference was only apparent when prediction models were built for fusiform rust resistance where loci of relatively larger effect had been described, although the increase in prediction accuracy was modest, from 0.29 by RR-BLUP to 0.34 by BayesA (Resende et al. 2012c). Considering the overall efficient performance of RR-BLUP for GS prediction in a number of studies for different species and traits, a general recommendation has been made to use it as a starting point from which to explore additional alternative models. These would include both Bayesian methods, when suspicion or prior information exists regarding the existence of some loci of larger effect, or machine learning methods when non-additive effects are known or presumed important (Lorenz et al. 2011; Heslot et al. 2012). Particularly in forest tree species where GS is intended not only for parent selection but also for clonal selection (e.g. *Eucalyptus*), analytical methods that can capture non-additive effects, should receive greater attention.

26.5.2 Will GS Be Able to Predict Non-Additive Genetic Effects?

GS has generally been discussed in the context of estimating breeding values. This probably comes from the fact that GS originally emerged as an approach in animal improvement where breeding strategies rarely exploit specific combining ability by mate allocation and do not involve the estimation of genotypic values of individuals to be deployed as clones. However for several plants species, vegetative propagation of outstanding individuals is an old and widely used method. Clones maximize gains from selection because all kinds of genetic effects, additive and non-additive, are captured. In several forest trees species, such as eucalypts and poplars, elite individuals are selected not only to serve as parents of the next breeding generation but also to be clonally propagated to provide superior planting material. For traits that display considerable levels of non-additivity top parents may not be top clones and vice-versa. GS applied to such species should therefore contemplate models to predict breeding values for parent selection or total genotypic values for clone selection.

The inclusion of non-additive effects into GS models was first considered in the context of bi-parental populations of crops (Lorenzana and Bernardo 2009) and mate allocation in animal breeding (Toro and Varona 2010). Recently, a stochastic simulation study of GS including dominance effects directed to *Eucalyptus* breeding was reported (Denis and Bouvet 2013). It showed that a GS model including dominance effects performed better for clone selection only when dominance effects were preponderant (i.e. a dominance to additive variance ratio approaching one) and heritability was > 0.6 . The inclusion of non-additive effects did not improve

the estimation of breeding value for parent selection. In *Eucalyptus*, particularly in hybrid breeding, non-additive effects tend to be significant. A dominance to additive variance ratio close to 1.2 for growth was estimated in a set of 684 full sib-families in 10 subpopulations from factorial crosses amongst 88 females and 107 males of *E. grandis* \times *E. urophylla* (Bouvet et al. 2009). In *E. globulus* a non-additive/additive variance ratio of 0.8 was estimated, with indications that epistasis might be the foremost component of the non-additive variance (Araujo et al. 2012). Initial studies involving the estimation of non-additive effects have been reported for crop plant situations, suggesting promising ways to incorporate epistasis into GS models (Xu and Jia 2007; Hu et al. 2011). Furthermore, the inclusion of over-dominance effects in prediction models should also be a topic of interest for the application of GS in predicting heterotic transgressive phenotypes, frequently observed and selected in inter-specific hybrid families of *Eucalyptus* (Assis 2011). At this point, no experimental data exist in forest trees regarding the ability of GS to predict the total genotypic value of individual trees including additive and non-additive effects. Research into this topic is thus seen as one of the top priorities for forest tree species that are deployed as clonal varieties.

26.5.3 What Will be the Accuracy of Predictions as Generations of GS Advance?

While GS can be reliably advocated for short-term gains, no such assurance is yet fully warranted for long-term gain, a crucial aspect for forest tree breeding. As pointed out earlier (Grattapaglia and Resende 2011), a critical question is: how accurate will the genomic predictions be on individuals several generations removed from the training population? Results from experimental studies so far, while very promising, are only based on cross validation within the same population. In operational GS, the selection candidates will rarely belong to the same population as the training set, and may well be several generations removed from it. As generations advance, recombination erodes marker-QTL LD reducing accuracy, while directional selection may change both the genetic architecture, via changes in allele frequencies, and the patterns of LD making it potentially unfavorable for GEBV prediction. In the seminal study of Meuwissen et al. (2001) the decline of GS accuracy over generations was estimated at 5 % per generation, getting smaller in later generations. Given the unfeasibility of experimental studies to assess the dynamics of long term GS over generations, its performance has not yet been examined experimentally, but several studies approached this issue by simulations, some of them with more complex models including the effect of directional selection (Muir 2007; Sonesson and Meuwissen 2009; Jannink 2010; Iwata et al. 2011; Long et al. 2011). All these studies converged to a fundamental recommendation: marker effects have to be re-estimated frequently in order to maintain accuracy of predictions over generations.

The issue of model updating was specifically assessed for a 60-year conifer tree breeding program by comparing the performance of GS with conventional phenotypic

selection using stochastic simulations (Iwata et al. 2011). Results showed that GS outperformed phenotypic selection in the short-term (30-years), but not in the long-term (60 years). When the prediction model was updated, however, the genetic gain of GS was nearly twice that of phenotypic selection, even for low heritability traits, with a greater advantage of GS as genotyping density increased. Two model updating strategies were tested: (1) a more conventional one where the prediction model generated in the initial cycle of selection is updated after three (or more) generations of GS by a conventional cycle of breeding where trees are grown and phenotyped to re-estimate the marker effects; and (2) an alternative strategy where in each cycle of GS an extra set of a few hundred progeny individuals of that cycle is actually planted in a field experiment. After a few years (depending on the species), phenotypes for that extra set of trees become available and are used to update the prediction model. From that point on, every year the prediction model gets updated with the inclusion of phenotypic data of the extra set of trees from previous generations. Because a set of trees from every cycle of GS is actually field grown, this second updating strategy allows continuous verification of the genetic progress of the GS program, although it involves greater costs of growing and measuring trees every generation and could theoretically increase the probability of unintended fixation of unfavorable alleles (Iwata et al. 2011). A significant advantage of model updating on GS accuracy by including phenotypic data from previous cycles was also shown by simulations in the context of *Eucalyptus* breeding (Denis and Bouvet 2013). From the practical standpoint of a breeding program, continuously associating phenotypic data from previous cycles of GS and thus progressively updating prediction models seems to be a very sensible and feasible approach to adopt. This extra set of a few hundred trees would have already been genotyped in the GS cycle anyway, and growing and measuring them would not represent a significant cost while allowing for permanent monitoring of the realized performance of GS.

Finally, two additional issues have been raised regarding the performance of GS over the long term: inbreeding and loss of useful variation. GS could potentially result in a fast and unintended frequency increase of deleterious alleles causing inbreeding depression or fixation of unfavorable QTL alleles due to the progressive restriction of population size. Daetwyler et al. (2007) have shown, however, that GS reduces the rate of inbreeding per generation when compared with sib and BLUP selection. High accuracies of estimated breeding values are achieved through better prediction of the Mendelian sampling term. This genomic-level resolution increases differentiation among sibs, allowing the breeder to better manage coancestry and to mitigate the rate of inbreeding even when selecting related individuals in breeding programs that are pushing for high genetic gains. Consistent with this expectation, the effect of non-random mating on the rate of inbreeding was recently found to be smaller for breeding schemes that adopt genome predictions when compared to conventional mating and selection designs (Nirea et al. 2012). Nevertheless Iwata et al. (2011) pointed out that periodic verification of performance of a subset of selected trees along the GS cycles of a breeding program is warranted to monitor any possible reduction of vigor attributable to weakly or moderately deleterious mutations.

The second issue regarding the impact of GS over time relates to the loss of favorable alleles with the faster successive cycles of breeding, potentially causing a progressive reduction of response to selection. Measures to mitigate this effect include using higher genotyping densities and periodical or continuous model updating, as discussed earlier. Additionally it has been shown that adopting weighed GS (Goddard 2009) together with using a larger training set (Jannink 2010), will help reducing the loss of low frequency favorable alleles in the breeding population, although some will inevitably be lost due to low LD with any genotyped marker. In a simulation study Jannink (2010) showed that placing additional weight on low-frequency favorable marker alleles, allowed GS to increase their frequency earlier on, causing an initial increase in genetic variance. This procedure led to higher long-term gain while mitigating losses in short-term gain. Weighted GS also increased the maintenance of marker polymorphism, ensuring that QTL-marker linkage disequilibrium was higher than in conventional unweighted GS.

26.5.4 Will GS Models Work Across Different Populations?

The maintenance of predictive ability of a GS model across different populations will essentially rely on the consistency of LD across populations which in turn depends on the recombination rate between marker and QTLs and the time since the two populations diverged. Therefore, the less diverged the populations are and the higher the marker density is, better performance of the predictive model is expected across populations. Multi-population training sets can be used which coupled to higher marker density, may capture the LD that existed prior to population divergence and provide good predictive ability. However if populations with different LD relationships are combined to train a GS model, markers may turn out not be predictive in any of them.

Breeding for inter-specific hybrids is an established strategy in some of the main plantation species. Hybrids combine desirable traits from two or more species through complementation of additive gene action (ex. growth and fungal disease resistance as in the *E. grandis* × *E. urophylla* hybrid) and heterotic effects due to non-additive gene action, frequently exhibiting greater phenotypic stability that allows extending plantation range to sites where one or both parental species have a suboptimal performance. Breeding for hybrids may be done by a (1) multi-population approach where each species is kept as a separate breeding population or a single population approach where the original species are hybridized at the outset to form a single hybrid breeding population. A third approach called reciprocal recurrent selection between synthetic multi species populations is becoming increasingly popular for *Eucalyptus* breeding in Brazil. Trait complementation and heterosis are exploited at the same time from up to six different species to improve drought, disease resistance, volume, density, yield and lignin content (Resende and de Assis 2008). In these cases GS models could be either trained by using individuals from both populations simultaneously to predict performance in F1 hybrids or trained in F1 hybrids to predict performance in the two separate populations. Studies on what could be

the best approach assuming variable genotyping densities and levels of divergence between the two species are needed. However the strong LD generated in each hybrid population should be favorable to GS.

An analogous situation takes place in bovine breeding in which selection is carried out in pure breeds but the aim is to improve crossbred performance. The issue of GS prediction across breeds or between crossbred and purebred populations has been extensively studied by simulations (de Roos et al. 2009; Ibanz-Escriche et al. 2009; Kizilkaya et al. 2010). Results generally show that crossbred data may provide adequate prediction accuracy for selecting purebred individuals for crossbred performance, although in all cases higher marker densities are required. Experimental data of GS across populations is still sparse however. Results in bovine breeds have shown that estimating marker effects within one breed and predicting performance of another breed results in low accuracies, although accuracies were nearly as good or better when using a combined-breed versus within-breed training population (Hayes et al. 2009a). Experimental GS data in two unrelated populations of *Eucalyptus* evaluated in two different environments showed that prediction models developed for one population could not predict phenotypes in the other. Results indicated that prediction models will in principle be population specific, although the experiment could not disentangle the confounding effect of genotype by environment interaction which may have played a greater role in explaining the loss of accuracy in validation across populations (Resende et al. 2012a).

26.5.5 Will GS Models Work Across Different Environments?

Genotype by environment ($G \times E$) interaction is a fact of life that all tree breeding programs commonly deal with, although at different levels, depending on the species, environmental variability and extent of the intended forest plantation sites and type of planting material, whether families or clones, with clones typically being more interactive than families. $G \times E$ is essentially a lack of consistency in the relative performance of individuals when they are grown in different environments. Interactions can be more subtle when differences in performance are observed but the relative ranking of tested individuals does not change across different environments (termed scale effect interaction), or more severe types of interactions when rank changes are observed. Correct ranking of individual trees by their GEBV or genotypic value is a key component of the successful implementation of GS. Therefore, while the presence of scale effect interactions should not represent a major limitation of a prediction model, rank changes are critical. When large rank change interactions are found, the GS strategy must account for this.

Considerations and treatment of the interaction between genome predictions and environment will essentially follow the same procedures used in dealing with standard $G \times E$ effects. Technically there is nothing different between dealing with conventional $G \times E$ or genomic effect by environment interaction. The same consideration regarding the definition of breeding or management zones (i.e. the set of

environments for which an improved variety is being developed), commonly applied to tree breeding programs, will likely apply to GS as well. Prediction models will be expected to be accurate across sites within the same breeding zone but not necessarily so across breeding zones. However the need to develop specific GS models for each breeding zone will largely depend upon the type of interaction observed, whether scale effect or rank change. In the only study so far in forest trees, the assessment of $G \times E$ in the context of genome predictions corroborated what was expected based on previous knowledge of $G \times E$ trends in loblolly pine along the southeastern USA (Resende et al. 2012b). The accuracies of models predicting GEBV were higher for the same site and declined for different sites. However the decline in accuracy between sites within the same breeding zone was only marginal when compared to the decline observed between distant breeding zones, suggesting that $G \times E$ more rigorously affect the transferability of models across breeding zones.

Most GS studies to date have dealt with a single-environment model, presumably because GS studies have been directed largely to animal breeding, where a common 'environment', i.e. animal management system, is generally assumed. Furthermore most estimates of $G \times E$ in controlled experiments of bovines were low to zero, while between-country studies reported only scaling effects and a few of them rank change ones (Hammami et al. 2009). In plants, however, multi-environmental $G \times E$ interaction not only is commonplace but it is widely used to assess the performance of the same genotypes (lines or clones) or families across different environmental conditions to study genotype stability and to predict the performance of untested genotypes. Studies of $G \times E$ in the context of GS for plant breeding are still rare, however. Heffner et al. (2009) pointed out that GS opens the opportunity to evaluate the effect of particular genomic segments that are shared between lines across multiple environments. With GS, lines will not be evaluated exclusively on the basis of their own phenotypic performance, but on the basis of information shared across other lines, years and environments. This information sharing should provide GS with stability of predictions even in the presence of $G \times E$. This same concept was recently put in practice by Burgueno et al. (2012) using a multi-environment dataset of wheat lines, showing that combining pedigree and marker data can yield substantial increases in prediction accuracy relative to traditional pedigree-based prediction and to single-environment pedigree and genomics prediction models. Multi-environment GS models enhanced predictive power in across-environment prediction, i.e. predicting the performance of genotypes that were evaluated in some environments but not in others, an application of significance in most plant breeding programs.

26.6 Conclusions

The prospects of GS applied to forest tree breeding are very promising. Both, simulation studies and experimental reports point in this direction. Despite biological and breeding system differences from those of forest trees, GS in dairy cattle breeding is a useful benchmark. Several countries have started the operational implementation of

GS programs for dairy cattle breeds. GEBV have been released and market share for genomically tested bulls is rapidly increasing, reaching 25–50% in some countries, while research into the best strategies and designs for optimizing a GS program are underway (Pryce and Daetwyler 2012). GS is currently a hot topic in plant breeding and crop genomics (Morrell et al. 2012) and several major crop and tree breeding programs worldwide are seriously looking into it. How to incorporate GS into a forest tree breeding program will vary on a case by case basis following a detailed cost/benefit analysis which may well conclude that it is not an option for now. It seems however that time gains by replacing progeny testing and streamlining clonal testing by GS will be inevitable, notwithstanding the major gain in allowing simultaneous evaluation of all traits in all progeny individuals in a single ‘genetic shot’.

An attempt was made here to discuss some of the main challenging issues that remain to be evaluated for the implementation of GS in forest tree breeding. These issues will likely receive increased attention from several research programs in the years to come. It is important to note, however, that some of the challenges and opportunities raised are valid under current technologies and circumstances. GS is a fast moving area of research and while some of the fundamental genetic issues discussed here are not likely to change much, some others will almost certainly change in the future as new genotyping and sequencing technologies emerge, paralleled by improved statistical approaches. In the meanwhile, breeding programs should increase efficiency by taking advantage of this modern breeding tool that GS represents, aware of the risks inherent to every innovation.

Genomic Selection has been anecdotically criticized by some for being a ‘black box’ method which retrocedes to what quantitative genetics has been for decades. By emphasizing phenotype prediction instead of trait dissection and gene discovery, no contribution is made in understanding the complex molecular mechanisms underlying quantitative genetic variation. A more practical ‘breeder’s’ attitude will not see any problem with that, based on the premise that a true and complete understanding of the molecular biology underlying complex traits might remain as an unachievable ‘holy grail’. In the meanwhile GS will allow breeders to make faster and possibly more precise genetic progress. After all, the infinitesimal model involving many genes with small effects appears to be nearly accurate for the traits assessed so far, and the methods developed for GS appear to work well in capturing most of these effects. Furthermore, notwithstanding the fact that discovering single quantitative trait nucleotides for complex traits represents a challenging task, even if a single one is discovered, it may not be useful in selection, either because its net effect on the selection index is negligible, or because the target allele is already found at high frequency in the population.

Alternatively, however, GS research and application and its prospective evolution can be viewed as a fresh way to provide a better paradigm and experimental framework, beyond QTL mapping, toward an improved connection between quantitative and molecular genetics. Recently a complementary approach exploiting large-scale genomic and metabolic information was used to predict complex, highly polygenic traits in maize hybrids (Riedelsheimer et al. 2012). The evolution of such integrative approaches based on huge genotype and phenotype datasets coupled to predictive

methods, could provide important additional hints through which to uncover connections between genome-wide elements such as microRNAs, retrotransposons, CNV and epigenetic signals, and the observable phenotypic variation in complex traits. Still, the elucidation of such connections will not be an easy task due to the time and space dynamics of the effects of these genomic elements, and the stochastic processes that thwart a one-to-one relation between genotype and phenotype. Supportive of the idea that the phenotypic era of plant breeding is irreplaceable and the challenge is enabling more effective phenotypic selection (Lee 2006), GS goes one step further in that direction by providing a more unbiased way to ‘place genotypes underneath a phenotype’ at the genome-wide scale.

Acknowledgments This work was supported by CNPq grant 577047/2008-6, PRONEX-FAP-DF grant “NEXTREE” 2009/00106-8, EMBRAPA Macroprogram 2 grant 02.07.01.004 and a CNPq research fellowship to DG.

References

- Albrecht T, Wimmer V, Auinger HJ, Erbe M et al (2011) Genome-based prediction of testcross values in maize. *Theor Appl Genet* 123:339–350
- Araujo JA, Borralho NMG, Dehon G (2012) The importance and type of non-additive genetic effects for growth in *Eucalyptus globulus*. *Tree Genet Genomes* 8:327–337
- Assis T (2011) Hybrids and mini-cutting: a powerful combination that has revolutionized the *Eucalyptus* clonal forestry. *BMC Proceedings* 5:118
- Assis TF, de Resende MDV (2011) Genetic improvement of forest tree species. *Crop Breed Appl Biotechnol* 11:44–49
- Baird NA, Etter PD, Atwood TS, Currey MC et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *Plos One* 3(10):e3376
- Bernardo R (2008) Molecular markers and selection for complex traits in plants: Learning from the last 20 years. *Crop Sci* 48:1649–1664
- Bernardo R, Yu JM (2007) Prospects for genomewide selection for quantitative traits in maize. *Crop Sci* 47:1082–1090
- Bouvet JM, Saya A, Vigneron P (2009) Trends in additive, dominance and environmental effects with age for growth traits in *Eucalyptus* hybrid populations. *Euphytica* 165:35–54
- Brondani RP, Williams ER, Brondani C, Grattapaglia D (2006) A microsatellite-based consensus linkage map for species of *Eucalyptus* and a novel set of 230 microsatellite markers for the genus. *BMC Plant Biol* 6:20
- Burgueno J, de los CG, Weigel K, Crossa J (2012) Genomic prediction of breeding values when modeling genotype x environment interaction using pedigree and dense molecular markers. *Crop Sci* 52:707–719
- Coster A, Bastiaansen JWM, Calus MPL, van Arendonk JAM et al (2010) Sensitivity of methods for estimating breeding values using genetic markers to the number of QTL and distribution of QTL variance. *Genet Sel Evol* 42:9
- Crossa J, de los CG, Perez P, Gianola D et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics* 186:713–406
- Daetwyler HD, Villanueva B, Bijma P, Woolliams JA (2007) Inbreeding in genome-wide selection. *J Anim Breed Genet* 124:369–376
- Daetwyler HD, Villanueva B, Woolliams JA (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One* 3:e3395

- Daetwyler HD, Hickey JM, Henshall JM, Dominik S et al (2010) Accuracy of estimated genomic breeding values for wool and meat traits in a multi-breed sheep population. *Anim Prod Sci* 50:1004–1010
- Denis M, Bouvet J-M (2013) Efficiency of genomic selection with models including dominance effect in the context of *Eucalyptus* breeding. *Tree Genet Genomes* 9:37–51
- Dillen S, Storme V, Marron N, Bastien C et al (2008) Genomic regions involved in productivity of two interspecific poplar families in Europe. 1. Stem height, circumference and volume. *Tree Genet Genomes* 5:147–164
- Echt CS, Saha S, Krutovsky KV, Wimalanathan K et al (2011) An annotated genetic map of loblolly pine based on microsatellite and cDNA markers. *BMC Genet* 12:17
- Elshire RJ, Glaubitz JC, Sun Q, Poland JA et al (2011) A robust, simple genotyping-by-sequencing (GbS) approach for high diversity species. *Plos One* 6:e19379
- Faria DA, Tanno P, Reis A, Martins A et al (2012) Genotyping-by-Sequencing (GbS) the highly heterozygous genome of *Eucalyptus* provides large numbers of high quality genome-wide SNPs. *Plant and Animal Genome Conference XX*, San Diego, p P0521
- Gion JM, Carouche A, Deweer S, Bedon F et al (2011) Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: *Eucalyptus*. *BMC Genomics* 12:301
- Goddard M (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* 136:245–257
- Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 10:381–391
- Grattapaglia D, Kirst M (2008) *Eucalyptus* applied genomics: from gene sequences to breeding tools. *New Phytol* 179:911–929
- Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. *Tree Genet Genomes* 7:241–255
- Grattapaglia D, Chaparro J, Wilcox P, Mccord S et al (1992) Mapping in woody plants with RAPD markers: applications to breeding in forestry and horticulture. *Proceedings of the Symposium “Applications of RAPD Technology to Plant Breeding”*. Crop Science Society of America, American Society of Horticultural Science, American Genetic Association, pp 37–40
- Grattapaglia D, Plomion C, Kirst M, Sederoff RR (2009) Genomics of growth traits in forest trees. *Curr Opin Plant Biol* 12:148–156
- Grattapaglia D, de Alencar S, Pappas G (2011) Genome-wide genotyping and SNP discovery by ultra-deep Restriction-Associated DNA (RAD) tag sequencing of pooled samples of *E. grandis* and *E. globulus*. *BMC Proceedings* 5:P45
- Greenwood MS, Adams GW, Gillespie M (1991) Stimulation of flowering by grafted black spruce and white spruce—a comparative-study of the effects of gibberellin a4/7, cultural treatments, and environment. *Can J For Res* 21:395–400
- Habier D, Fernando RL, Dekkers JCM (2009) Genomic Selection using low-density marker panels. *Genetics* 182:343–353
- Haley CS, Visscher PM (1998) Strategies to utilize marker-quantitative trait loci associations. *J Dairy Sci* 81:85–97
- Hammami H, Rekik B, Gengler N (2009) Genotype by environment interaction in dairy cattle. *Biotechnol Agron Soc* 13:155–164
- Harfouche A, Meilan R, Kirst M, Morgante M et al (2012) Accelerating the domestication of forest trees in a changing world. *Trends Plant Sci* 17:64–72
- Hasan O, Reid JB (1995) Reduction of generation time in *Eucalyptus globulus*. *Plant Growth Regul* 17:53–60
- Hayes B, Goddard M (2010) Genome-wide association and genomic selection in animal breeding. *Genome* 53:876–883
- Hayes BJ, Bowman PJ, Chamberlain AC, Verbyla K et al (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genet Sel Evol* 41:51
- Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009b) Invited review: genomic selection in dairy cattle: progress and challenges. *J Dairy Sci* 92:433–443

- Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement. *Crop Sci* 49:1–12
- Heslot N, Yang HP, Sorrells ME, Jannink JL (2012) Genomic selection in plant breeding: a comparison of models. *Crop Sci* 52:146–160
- Hu ZQ, Li YG, Song XH, Han YP et al (2011) Genomic value prediction for quantitative traits under the epistatic model. *BMC Genet* 12:15
- Hudson CJ, Freeman JS, Kullán AR, Petroli CD et al (2012) A reference linkage map for *Eucalyptus*. *BMC Genomics* 13:240
- Ibanz-Escriche N, Fernando RL, Toosi A, Dekkers JCM (2009) Genomic selection of purebreds for crossbred performance. *Genet Sel Evol* 41:12
- Iwata H, Hayashi T, Tsumura Y (2011) Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. *Tree Genet Genomes* 7:747–758
- Jannink JL (2010) Dynamics of long-term genomic selection. *Genet Sel Evol* 42:35
- Jannink JL, Zhong SQ, Dekkers JCM, Fernando RL (2009) Factors affecting accuracy from Genomic Selection in populations derived from multiple inbred lines: a barley case study. *Genetics* 182:355–364
- Jannink JL, Lorenz AJ, Iwata H (2010) Genomic selection in plant breeding: from theory to practice. *Brief Funct Genomics* 9:166–177
- Kizilkaya K, Fernando RL, Garrick DJ (2010) Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J Anim Sci* 88:544–551
- de Koning DJ, McIntyre L (2012) Setting the standard: a special focus on Genomic Selection in GENETICS and G3. *Genetics* 190:1151–1152
- Kumar S, Bink MCAM, Volz RK, et al (2012a) Towards genomic selection in apple (*Malus domestica* Borkh.) breeding programmes: Prospects, challenges and strategies. *Tree Genet Genomes* 8:1–14
- Kumar S, Chagne D, Bink MCAM, Volz RK et al (2012b) Genomic Selection for fruit quality traits in apple (*Malus domestica* Borkh.). *PLoS One* 7:e36674
- Lee M (2006) The phenotypic and genotypic eras of plant breeding. In: Lamkey KR, Lee M (eds) *Plant breeding: the Arnel R Hallauer international symposium* Blackwell Publishing, Ames, pp 213–217
- Legarra A, Robert-Granie C, Manfredi E, Elsen JM (2008) Performance of genomic selection in mice. *Genetics* 180:611–618
- Long N, Gianola D, Rosa GJM, Weigel KA (2011) Long-term impacts of genome-enabled selection. *J Appl Genet* 52:467–480
- Lorenzana RE, Bernardo R (2009) Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theor Appl Genet* 120:151–161
- Lorenz AJ, Chao SM, Asoro FG, Heffner EL et al (2011) Genomic Selection in Plant Breeding: Knowledge and Prospects. *Adv Agron* 110(110):77–123
- Luan T, Woolliams JA, Lien S, Kent M et al (2009) The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics* 183:1119–1126
- McKeand SE, Bridgwater FE (1998) A strategy for the third breeding cycle of loblolly pine in the Southeastern US. *Silvae Genetica* 47:223–234
- Meilan R (1997) Floral induction in woody angiosperms. *New For* 14:179–202
- Meuwissen TH, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. *Nat Rev Genet* 13:85–96
- Moser G, Tier B, Crump RE et al (2009) A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet Sel Evol* 41:56
- Muir WM (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet* 124:342–355

- Myles S, Chia JM, Hurwitz B, Simon C et al (2010) Rapid genomic characterization of the genus *Vitis*. *Plos One* 5:e8219
- Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? *Ann Bot (Lond)* 110:1303–1316
- Namkoong G, Kang HC, Brouard JS (1988) *Tree Breeding: principles and strategies*. Springer Verlag, New York
- Neale DB, Williams CG (1991) Restriction-Fragment-Length-Polymorphism mapping in conifers and applications to forest genetics and tree improvement. *Can J For Res* 21:545–554
- Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. *Nat Rev Genet* 12:111–122
- Nielsen HM, Sonesson AK, Yazdi H, Meuwissen THE (2009) Comparison of accuracy of genome-wide and BLUP breeding value estimates in sib based aquaculture breeding schemes. *Aquaculture* 289:259–264
- Nirea KG, Sonesson AK, Woolliams JA, Meuwissen THE (2012) Effect of non-random mating on genomic and BLUP selection schemes. *Genet Sel Evol* 44:11
- Novaes E, Osorio L, Drost DR, Miles BL, Boaventura-Novaes CRD et al (2009) Quantitative genetic analysis of biomass and wood chemistry of *Populus* under different nitrogen levels. *New Phytol* 182:878–890
- Poland JA, Brown PJ, Sorrells ME, Jannink JL (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7:e32253
- Pryce JE, Daetwyler HD (2012) Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Anim Prod Sci* 52:107–114
- Rae A, Pinel M, Bastien C, Sabatti M et al (2008) QTL for yield in bioenergy *Populus*: identifying G × E interactions from growth at three contrasting sites. *Tree Genet Genomes* 4:97–112
- Raymond CA, Schimleck LR (2002) Development of near infrared reflectance analysis calibrations for estimating genetic parameters for cellulose content in *Eucalyptus globulus*. *Can J For Res* 32:170–176
- Resende MDV, de Assis TF (2008) Seleção recorrente recíproca entre populações sintéticas multi-espécies (SRR-PSME) de eucalipto. *Pesquisa Florestal Brasileira* 57:57–60
- Resende MDV, Resende MFR, Sansaloni CP, Petrolí CD et al (2012a) Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. *New Phytol* 194:116–128
- de Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. *Genetics* 183:1545–1553
- Resende MFR, Munoz P, Acosta JJ, Peter GF et al (2012b) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. *New Phytol* 193:617–624
- Resende MFR, Munoz P, Resende MDV, Garrick DJ et al (2012c) Accuracy of genomic selection methods in a standard data set of Loblolly Pine (*Pinus taeda* L.). *Genetics* 190:1503–1510
- Riedelshheimer C, Czedik-Eysenberg A, Grieder C, Lisek J et al (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat Genet* 44:217–220
- Sansaloni C, Petrolí C, Jaccoud D, Carling J et al (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proceedings* 5:P54
- Sewell MM, Sherman BK, Neale DB (1999) A consensus map for loblolly pine (*Pinus taeda* L.). I. Construction and integration of individual linkage maps from two outbred three-generation pedigrees. *Genetics* 151:321–330
- Silva JCE, Wellendorf H, Borralho NMG (2000) Prediction of breeding values and expected genetic gains in diameter growth, wood density and spiral grain from parental selection in *Picea abies* (L.) KARST. *Silvae Genetica* 49:102–109
- Solberg TR, Sonesson AK, Woolliams JA et al (2009) Persistence of accuracy of genome-wide breeding values over generations when including a polygenic effect. *Genet Sel Evol* 41:53

- Sonesson AK, Meuwissen THE (2009) Testing strategies for genomic selection in aquaculture breeding programs. *Genet Sel Evol* 41:37
- Strauss SH, Lande R, Namkoong G (1992) Limitations of molecular-marker-aided selection in forest tree breeding. *Can J For Res* 22:1050–1061
- Sved JA (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theor Popul Biol* 2:125–141
- Thumma BR, Southerton SG, Bell JC, Owen JV et al (2010) Quantitative trait locus (QTL) analysis of wood quality traits in *Eucalyptus nitens*. *Tree Genet Genomes* 6:305–317
- Toro MA, Varona L (2010) A note on mate allocation for dominance handling in genomic selection. *Genet Sel Evol* 42:33
- Van Tassell CP, Smith TPL, Matukumalli LK, Taylor JF et al (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* 5:247–252
- VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS et al (2009) Invited review: Reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* 92:16–24
- White TL, Adams WT, Neale DB (2007) *Forest Genetics*. CABI Publishing
- Williams CG (1988) Accelerated short-term genetic testing for Loblolly Pine families. *Can J For Res* 18:1085–1089
- Williams CG, Neale DB (1992) Conifer wood quality and marker-aided selection—a case-study. *Can J For Res* 22:1009–1017
- Wright S (1931) Evolution in Mendelian populations. *Genetics* 16:97–159
- Xu SZ, Jia ZY (2007) Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics* 175:1955–1963
- Zapata-Valenzuela J, Isik F, Maltecca C, Wegrzyn J et al (2012) SNP markers trace familial linkages in a cloned population of *Pinus taeda* – prospects for genomic selection. *Tree Genet Genomes* 8:1307–1318
- Zhao YS, Gowda M, Liu WX, Wurschum T et al (2012) Accuracy of genomic selection in European maize elite breeding populations. *Theor Appl Genet* 124:769–776

Chapter 27

Genetic Diversity in the Grapevine Germplasm

Federica Cattonaro, Raffaele Testolin, Simone Scalabrin, Michele Morgante
and Gabriele Di Gaspero

Contents

27.1	Introduction	684
27.2	Grapevine Taxa and Botanical Context	685
27.2.1	Taxonomy	685
27.2.2	Wild Grapevines	686
27.2.3	Origin of <i>Vitis vinifera</i>	687
27.3	Domesticated Grapevines and Their Wild Progenitor(s)	688
27.3.1	Change in Mating System	688
27.3.2	Tempo and mode of domestication	689
27.4	Bi-directional Gene Flow Between Vineyards and the Wilderness	690
27.5	Ancient Cultivars	691
27.6	Selective Sweeps	693
27.7	The Effect of Modern Breeding	693
27.8	Genetic Diversity in Traditional Cultivars	694
27.8.1	Global Patterns of Genetic Structure	694
27.8.2	The Classification of Negrul into Proles	695
27.8.3	Kinship and Founders of Major Descent Groups	696
27.8.4	Heterozygosity	697
27.8.5	DNA Sequence Diversity	698
27.8.6	Haplotype Diversity	699
27.8.7	Patterns of Linkage Disequilibrium	700
27.8.8	Structural Diversity	700
	References	701

Abstract Grapevine is a major horticulture crop grown on ~7.6 million hectares that secure a yearly production of ~70 million tons of grapes. A significant part of the crop (65 %) annually fuels a worldwide wine industry of ~27 billion liters. In 2007,

G. Di Gaspero (✉) · F. Cattonaro · R. Testolin · S. Scalabrin · M. Morgante
Istituto di Genomica Applicata, Parco Scientifico e Tecnologico Luigi Danieli,
33100 Udine, Italy
e-mail: gabriele.digaspero@uniud.it

G. Di Gaspero · R. Testolin · M. Morgante
Dipartimento di Scienze Agrarie e Ambientali, University of Udine,
via delle scienze 208, 33100 Udine, Italy

2.8 billion liters in wine sales in the United States alone were worth US\$ 30 billion. Grapevines and wine making are also an integral part of the landscape and the cultural heritage of Southern Europe and Southwestern Asia, in the native environments of the species *Vitis vinifera*. The natural diversity that existed in grapevine was key to its successful colonization of the Mediterranean shores and continental Europe, the home to the modern wine industry and the main source for the dispersal of domesticated grapevines to other continents along trade routes. The sequencing of the grapevine genome has attracted renewed interest in this ancient crop, which largely relies on centuries-old varieties immortalized by vegetative propagation, and it has revived the exploration of germplasm and the analysis of genetic diversity, which are now afforded with the application of powerful technologies of DNA analysis.

Keywords *Vitis vinifera* · Vitaceae · Selective sweeps · Breeding · Descent groups

Abbreviations

AD	<i>Anno Domini</i>
BC	Before Christ
cpDNA	chloroplast DNA
cM	centiMorgan
EAS-ENA	Eastern Asia—Eastern North America
LD	Linkage disequilibrium
MYA	Million years ago
SNP	Single nucleotide polymorphisms

27.1 Introduction

The genetic diversity in present-day grapevines reflects a combination of bottlenecks, drifts, and selection. The most significant events include early radiation of the Vitaceae ancestor from other rosids some 112–101 million years ago (MYA), diversification within the genus *Vitis* during the past 6 MYA (Zecca et al. 2011), and domestication of *Vitis vinifera* in Southwestern Asia from one or more East Asian progenitors between the 7th and the 4th millennia BC. Since then, anthropogenic processes such as human migration, trade, and land conquests dispersed the domesticated grapevines in a highly fragmented manner, which established sympatry between allochthonous domesticated forms and local wild forms of *V. vinifera* subsp. *sylvestris*. Admixture led the introduced accessions to hybridize to one another, and established a bi-directional gene flow between allochthonous domesticated forms and autochthonous *sylvestris*. Some uncertainty remains as to whether the autochthonous wild forms in Western Europe only contributed to present-day western varieties through gene introgression or some western varieties were domesticated *ex novo* from western *sylvestris*. Understanding the evolutionary changes from antiquity to modern times that gave rise to the diversity of grapevines, in particular those

associated with the domestication syndrome (e.g. hermaphroditism, survival amidst major pathogens, size, color, and composition of the berry), is the greatest challenge to discerning the relevance of past events that have shaped the cultivated germplasm for ages.

27.2 Grapevine Taxa and Botanical Context

27.2.1 Taxonomy

All grapevines belong to the family of Vitaceae. The phylogenetic placement of the Vitaceae remains equivocal. This family is tentatively placed as sister to the remaining rosids (Jansen et al. 2006), but it could also be placed within the rosids to form a clade with the Saxifragales, from which it has diverged since an estimated 112–101 MYA (Moore et al. 2010), or it could be placed as a sister to the rest of Rosidae (Soltis et al. 2011). Despite this controversy, all evidence is consistent with early divergence of Vitaceae during the evolution of Pentapetalae. Vitaceae are normally lianas with climbing ability provided by modified inflorescences that develop into tendrils opposite the leaves. Striking exceptions are succulent species in the genera *Cissus* and *Cyphostemma*, which consist of highly diverse plants and include the highest number of species (367 and 258, respectively). Vitaceae bear edible fleshy fruits containing 1–4-seeds. The family is mostly pantropical in the range of distribution, with a few genera (e.g. *Ampelopsis*, *Muscadinia*, *Parthenocissus*, and *Vitis*) exclusively indigenous to temperate regions of the Northern Hemisphere, some of them displaying an Eastern Asia—Eastern North America (EAS-ENA) disjunction pattern. Taxa of agricultural interest are found in the genera *Vitis* and *Muscadinia*. The genus *Vitis* includes an estimated 68 species, while only 3 species are classified in the genus *Muscadinia*. There are also species of ornamental merit in the genera *Ampelopsis* and *Parthenocissus*, which are popular climbers in temperate climate gardens, such as the Boston-ivy *P. tricuspidata* of China and Japan, the North American Virginia creeper *P. quinquefolia*, the Asian porcelain berry *A. brevipedunculata*, and to a lesser extent *A. aconitifolia* and *A. macrophylla*. The kangaroo vine *Cissus antarctica* and the Rex Begonia vine *Cissus discolor* are ornamental climbers used in subtropical climates.

Karyotyping revealed an extensive variation in chromosome number among the 14 genera of Vitaceae, ranging from $2n = 22$ in species of *Tetrastigma* to extremes of $2n > 80$. *Ampelocissus*, *Ampelopsis*, *Clematicissus*, *Parthenocissus*, and *Muscadinia* are generally $2n = 40$, while all *Vitis* species are $2n = 38$. All attempts to hybridize *Vitis* species with the closely related *Ampelopsis* and *Parthenocissus* have proven unsuccessful. In contrast, karyotypic differences do not prevent *Vitis* × *Muscadinia* crosses from generating viable progeny, provided that *Vitis* is used as seed parent and *Muscadinia* as pollen donor (Bouquet 1980). Karyotyping has provided clues that species in *Vitis* and *Muscadinia* are ancient polyploids with three basic sets of chromosomes, present in the combination $(6 + 7) + 6 = 19$ in *Vitis* or $(6 + 7) + 7 = 20$ in *Muscadinia*. The paleohexaploidy of *V. vinifera* has been confirmed by evidence of gene synteny across chromosome triplets (Jaillon et al.

2007). These two genera have only two sets of chromosomes in common, *Vitis* = 13 R^VR^V + 6AA and *Muscadinia* = 13 R^MR^M + 7BB (Patel and Olmo 1955). Thus, F₁ hybrids have 39 somatic chromosomes (13 R^VR^M + 6A + 7B), and most of them are sterile due to incorrect chromosome pairing and unbalanced separation during meiosis. On average, two to nine univalents are usually observed, with rare multivalents (Bouquet 1980). Occasionally, correct chromosomal pairing restored fertility in partially fertile hybrids (> 0.1 % ovule set), permitting genetic recombination to occur between chromosome complements (Jelenković and Olmo 1968). The few fertile *V. vinifera* × *M. rotundifolia* hybrids produced by Detjen (1919), Bouquet (1980), and Olmo (1986) provided the foundation for the process of introgression of disease resistance genes from *M. rotundifolia*—e.g. the powdery mildew resistance gene *Run1* (Pauquet et al. 2001)—and generated a descent group of remarkable size (Riaz et al. 2008).

27.2.2 Wild Grapevines

Wild populations of grapevines are taxonomically classified into ecospecies. Sympatry across vast geographical areas in Asia and North America and the absence of fertility barriers have prevented the confinement of natural diversity into separate genetic pools, and have blurred boundaries between species (Péros et al. 2010; Tröndle et al. 2010). The ease with which grapevines naturally hybridized prompted Levadoux and coworkers (1962) to promote the concept that the genus *Vitis* is a unique gene pool. Clinal variation within ecospecies also obstructs the definition of the exact number of ecospecies, as there are several uncertain taxa in East Asia and North America. Ecospecies have distinctive morphological traits, habitats and phylogeographical history. Chloroplast and nuclear DNA suggested that the genus *Vitis* is monophyletic and sister to *Muscadinia*, the Asian *Vitis* are ancestral, and the EAS-ENA continental disjunction is recent. Two dispersal events have introduced two distinct Asian haplotypes into North America. The rarest haplotype is found in *V. californica* across the North American West Coast, while the other haplotype is shared by all other North American taxa. Within the latter lineage, rapid radiation has occurred in the East Coast, presumably since the Pleistocene (11.5–2.6 MYA), giving rise to many ecospecies with considerable diversity (Péros et al. 2010). Paleobotany confirmed the affinity between North American and Asian species. Fossil pips recovered from the Gray Fossil Site (7–4.5 MYA) in Tennessee, USA, were attributed to three grape fossil species, one of which shows similarity with the extant North American *V. labrusca* and the others resemble the modern Asian species *V. thumbergii* and *V. lanata* (Gong et al. 2010). Morphological commonalities in grapevine disjuncts, such as those observed between *V. labrusca* and the Korean-Japanese *V. coignetiae*, have also been explained by convergent evolution driven by independent adaptation to similar habitats. Divergence in nuclear DNA is lower among Asian taxa than among North American taxa, in spite of the fact that Asian species have diverged for a longer period of time in the centre of diversification of the genus.

Homogenization of Asian taxa may have occurred through gene flow in restricted forests along the present-day Chinese, Korean, and Japanese coastal areas, where grapevines and other temperate perennials retreated during the last glacial period. The relevance of wild species for the improvement of the crop has become evident since modern viticulture was faced with new pests and pathogens, and the range of cultivation of grapevines was broadened beyond the limits of natural distribution into more extreme environments. North American species have historically been used as a source of genetic variation for disease resistance and frost hardiness (Munson 1909), and many more species are proving equally useful as we gain a better understanding of the diversity present in Asian germplasm (Wan et al. 2008a, b).

27.2.3 *Origin of Vitis vinifera*

The lineages that gave rise to *V. vinifera* were dispersed from Far Eastern Asia to western Eurasia. Chloroplast DNA (cpDNA) haplotypes of *V. vinifera* in western Eurasia are compatible with the theory that there were at least two distinct progenitors from the Far East (Péros et al. 2010). The most ancestral *Vitis* cpDNA haplotype is shared by the species *V. coignetiae*, *V. flexuosa*, *V. piasezkii*, and *V. thumbergii* in Eastern Asia and several Mediterranean varieties of *V. vinifera*. This haplotype has diversified after the westward colonization of the ancestor into three derived haplotypes: one is present in cultivars such as Pinot Noir and Grenache, another is present in Cinsaut and Folle Blanche, and the third is found in Coarna Negra and Kefessia. A fourth *V. vinifera* haplotype is present, e.g. in Sauvignon and Chenin Blanc, and differs substantially from other chloroplast lineages as if it was donated by a divergent founder. Fossil remains indicate that *Vitis* and *Muscadinia* ancestors were present in Western Eurasia since ~55 MYA (Fairon-Demaret and Smith 2002). Later, *Muscadinia* and possibly other *Vitis* lineages, became extinct from Europe during the Quaternary glacial periods. The most ancient fossil pips of *V. vinifera* in Northern Europe date back to the Holstein Interglacial of the Pleistocene (Turner 1968). During the Ice Age, southward displacement of grapevines was obstructed in Western Eurasia by east-west mountain ranges (Pyrenees, Alps, Carpathians, Caucasus, and the Himalayas), contrary to the scenario in North America and Eastern Asia (Jackson 2000). Diversity suffered in the Old World, and survival of *V. vinifera* was restricted to warmer refuges in coastal and insular areas of the northern and southern side of the Mediterranean Basin, around the Black Sea, around the southern part of the Caspian Sea, and inland along the Danube, Rhein, and Rhône rivers. Glacial displacement shaped the current distribution of *V. vinifera* subsp. *sylvestris*, which covers a vast geographical range from Portugal to the foothills of the Himalayas in small disjoint populations. Retraction of *V. vinifera* subsp. *sylvestris* into riparian locations with anoxic soils and to gravely/sandy soils along the Mediterranean shores – both unsuitable for the development of the root-feeding phylloxera—saved relictic populations of native wild grapevines from the infestation of the North American introduced pest, which threatened the survival of cultivated varieties in most vineyard soils since the mid-1800s (Ocete et al. 2011).

27.3 Domesticated Grapevines and Their Wild Progenitor(s)

The process of grapevine domestication was associated with little modification in the architecture and biology of the grape plant in comparison with other crops (This et al. 2006). This can be explained in part by the popular claim that the founders of the most ancient cultivars are only a few generations removed from their *V. vinifera* subsp. *sylvestris* progenitors and by the vegetative propagation of selected varieties since the second half of the last millennium. Alternatively, this small phenotypic gap between cultivated varieties and their wild relatives might be due to independent domestication events that occurred more recently after the establishment of viticulture in Western Europe and/or to the continuous gene flow between *sylvestris* and *sativa* that homogenized the two subspecies. The most significant modifications associated with domestication are the development of hermaphroditic flowers in *V. vinifera* subsp. *sativa*, the increase in number of berries per cluster (fruitfulness), the enlargement of berry size, the seedlessness in table grapes generally through stenospermocarp and only exceptionally via parthenocarp, and the change in seed shape in seeded varieties. Seed morphology is the most conspicuous and stable character for distinguishing wild and cultivated forms, although arguments are debated for explaining the selective advantage, if any, associated with this modification (Terral et al. 2010). Seeds are spherically shaped with a small beak in wild forms, while cultivated varieties have pyriform seeds with longer beaks. An elongated and thinner shape suggests less-developed embryos and fewer reserves in seeds of cultivated varieties, which should lead to poor germination. It remains unclear whether the change in seed shape has any biological significance, or if it simply reflects the pleiotropic effect of a gene controlling another agronomic trait that was target of selection during domestication.

27.3.1 Change in Mating System

Hermaphroditism is widespread in Vitaceae, and commonplace in the genera *Ampelopsis* and *Parthenocissus*, but not in *Vitis* and *Muscadinia*, which have functionally dioecious plants, with the exception of the domesticated form of *V. vinifera*. Monoecy is therefore considered ancestral to dioecy in Vitaceae, and domestication in *V. vinifera* restored the original status. This belief is corroborated by the observation that all flower types in *Vitis* have a full complement of female and male structures. Functionally unisexual flowers eventuate from incomplete development of ovules or pollen grains within morphologically hermaphroditic flowers. In dioecious species of *Vitis* and *Muscadinia*, femaleness is associated with inaperturate pollen grains that fail to germinate. Male flowers have a rudimentary pistil with inconspicuous style and stigma, but they may also develop to the point that ovules, style, and stigma are well formed, and abnormality is restricted to incomplete development of the embryo sac. Floral dimorphism is also subtle in the unisexual flowers of female or male plants of *V. vinifera* subsp. *sylvestris*, which follow a functionally hermaphroditic development past the completion of gametogenesis. Sexual differentiation is functionally

established only once the degeneration of external cells causes the nucellus of ovules to detach from integuments in male flowers, and failure in forming colpi causes the microspores of female flowers to remain coated with a uniformly thickened wall (Caporali et al. 2003; Gallardo et al. 2009). Female varieties of *V. vinifera* subsp. *sativa* with inaperturate pollen are common, in particular in the *proles orientalis* with as many as 378 listed in the Vitis International Variety Catalogue (www.vivc.de). Many hermaphroditic cultivars are heterozygous and give rise to progeny with female flowered plants. Female varieties may also descend from poorly domesticated material or accidental introgression of dioecy from the wild. Pollen grains are only rarely fertile in female varieties of *V. vinifera* subsp. *sativa*. DNA genotyping of seedlings from female seed parents revealed that progeny from self-fertilized flowers were produced at a rate of 0.08%, when intact female flowers had been given fertile pollen from a different variety past flower opening (authors, unpublished data). Offspring that were raised from self-fertilization of supposedly female grapevines were also found in *V. vinifera* subsp. *sylvestris* (Di Vecchi-Staraz et al. 2009). *V. vinifera* subsp. *sylvestris* male vines also revert and partially function as hermaphrodites (Negi and Olmo 1970). The vast majority (80–97%) of vegetative replicates of an andromonoecious accession of *V. vinifera* subsp. *sylvestris* developed some hermaphroditic flowers that were capable of setting seeded fruit at rates rarely exceeding 1% (fruit set over 100 flower buttons), but in most instances was lower by one or two orders of magnitude. Exogenous application of cytokinins converts male flowers into hermaphroditic flowers (Negi and Olmo 1966). All of this evidence indicates that dioecy is superficial in *V. vinifera*, and individuals with self-pollinating flowers might have originated in wild populations several times by independent events of mutation/recombination at sex genes or at minor modifying genes (Negi and Olmo 1971), which would fit the bipartite gene model of dioecy described by Westergaard (1958).

27.3.2 *Tempo and mode of domestication*

Archaeological, historical, and genetic pieces of evidence indicate Southern Caucasia—which includes present-day Georgia, Azerbaijan, Northwestern Turkey, Armenia, and Northeastern Iran—as the most likely place of the earliest domestication of *V. vinifera* in the Late Neolithic, corresponding with the most ancient evidence of winemaking (Vavilov 1951; Olmo 1996; Zohary 1996; Miller 2008). Dispersal of domesticated grapevines is believed to have initially occurred through transportation of seeds by migrating human populations. Archaeological pips of the domesticated form of grapevine were excavated from Copper and Bronze Age sites in the Near East, Middle East, Asia Minor, Greece, Crete, and Cyprus. Domesticated grapevines were later shipped through the Mediterranean as viticulture expanded into the Italian and Iberian peninsulas, Southern France, and Maghreb approximately 3 millennia before the present time. From these coastal regions, European viticulture was progressively

established inland. The living wild progenitor, *V. vinifera* subsp. *sylvestris*, grew naturally in many Mediterranean and continental European locations near the vineyards where the domesticated varieties imported from Southwestern Asia were introduced. This sympatry raises the question as to whether modern local varieties have purely descended from domesticated founders, or if they are hybrid populations originating from gene exchange between domesticated and wild forms, or if they were selected from autochthonous wild forms through independent events of domestication. The contribution of genes/alleles from European and North African *V. vinifera* subsp. *sylvestris* to modern varieties has been supported by indisputable molecular evidence (El Oualkadi et al. 2011; Myles et al. 2011), while the way this contribution has occurred—through secondary centres of domestication or introgression from *V. vinifera* subsp. *sylvestris* — has not been firmly established.

27.4 Bi-directional Gene Flow Between Vineyards and the Wilderness

Domesticated grapevines are able to hybridize easily with their sympatric wild relatives. Pollen-mediated gene introgression has been bi-directional, enriching the introduced varieties with genes for local adaptation and eroding genetic diversity in wild forms. In both ways, introgression has reduced the divergence between cultivated and wild compartments. *V. vinifera* is predominantly wind-pollinated over short distances in vineyards, but Coleoptera, Hymenoptera, and Diptera are habitual visitors of grapevine flowers in wild conditions, ensuring long-range pollen dispersal (Brantjes 1978). Pollen flow between compartments may occur when wild forms grow naturally in the vicinity of vineyards, and when seedlings that arise from dispersed seeds from vineyards colonize the habitat of *V. vinifera* subsp. *sylvestris*. Wild forms exist as cohorts of a few individuals, and the crop-to-wild direction of pollen flow is expected to overwhelm the flow in the opposite direction, causing cryptic introgression and progressive genetic erosion of local forms (Zecca et al. 2010). Cross-pollination has been estimated to occur at a rate of 4–26 %, depending on the distance between the site of the wild populations and the closest vineyard (Di Vecchi-Staraz et al. 2009). Parentage analysis demonstrated that today seemingly wild individuals are in some cases feral grapevines—offspring of scion varieties or rootstocks once grown in the same area, which escaped from vineyards and mated in the wilderness. The presence of invasive naturalised populations in the ecological niche of *V. vinifera* subsp. *sylvestris* is a factor that threatens the survival of purely native wild grapevines and blurs data of population structure (Arrigo and Arnold 2007). Crop-to-wild pollen flow could have generated many hybrid cultivars in the past. Rhein Riesling is a variety distinguished by the petrol flavour imparted by trimethyl-dihydronaphthalene. Riesling has undergone introgression from western wild grapevines, as it originated during the Middle Ages somewhere along the Rhine River Valley in Germany from the hybridization of Gouais Blanc (synonymous with Heunisch Weiss) and an alleged seedling of *V. vinifera* subsp. *sylvestris* × Traminer.

Traminer (synonymous with Savagnin Blanc) itself has many morphological traits that resemble the forms of *V. vinifera* subsp. *sylvestris* in the riparian forests along the Rhine River (Barth et al. 2009), with the notable exceptions of monoecy and lack of anthocyanin pigmentation in berry skin. The absence of white-skinned grapes in the populations of *V. vinifera* subsp. *sylvestris* and the recessive nature of the white-skin mutations in the domesticated compartment (Walker et al. 2007) argue against the possibility that ancient white cultivars are first generation hybrids of purely wild grapevines. This argument is corroborated by indices of genetic diversity between old varieties (e.g. Pinot Noir, Traminer, Riesling) and single indigenous wild individuals in the Rhein and Danube River valleys (Perret et al. 2000).

27.5 Ancient Cultivars

The most ancient grape varieties have been historically documented in Western Europe since the Middle Ages, but the general belief claims that some cultivars may be much older. Pinot was first mentioned at the end of the fourteenth century, but to many historians, the description made by Columella of the variety that Roman conquerors encountered in Burgundy in the first century AD recalled distinctive features of Pinot Noir. Gouais Blanc was also dispersed north of the Alps since antiquity (Boursiquot et al. 2004). Historical records linked Gouais Blanc to the grape variety that the Roman Emperor Probus donated from his homeland Dalmatia (present-day Croatia) to the Gauls. Many modern varieties are offspring of Pinot and Gouais Blanc. The fact that many seedlings have arisen from the hybridization of Pinot and Gouais Blanc at different times and in different places, where the most valuable became new selected varieties, indicates that Pinot and Gouais Blanc have been widely cultivated for a very long time. Their presence north of the Alps necessarily predated the earliest reference to Chardonnay, the most famous among their offspring, which was first mentioned in 1330. By that time, Gouais Blanc should also have populated the Rhein River Valley, where it gave rise to the seedling that became Riesling. Based on historical records, Traminer is believed to have originated on the northern side of the Alps somewhere between southern Germany, where it was mentioned at the end of the fifteenth century, and France. Traminer was widely cultivated in ancient times and hybridized with local varieties in Austria, Germany, and northern and southern France, disseminating progeny that are known as old varieties (e.g. Silvaner, Grüner Veltliner, Sauvignon Blanc, Chenin Blanc, and Petit Manseng). Where Traminer originated remains a mystery. DNA profiles show close kinship between Traminer and two individuals of *V. vinifera* subsp. *sylvestris* (Regner et al. 2000), although the direction of introgression is not conclusively demonstrated. Pinot and Traminer share a parent-offspring relationship (Regner et al. 2000), but it is not known which predated the other. Small-sized bunches, round shaped leaves, and blistering of the leaf lamina in Pinot and Traminer are wild traits that are unusual in other domesticated forms, corroborating the hypothesis that their pedigree included a *V. vinifera* subsp. *sylvestris* ancestor within a few generations. Chasselas is another European variety that has been mentioned since the 16th-seventeenth century

in historical records across an area ranging from southern Germany to Switzerland, as well as the French Bourgogne. Speculation on its origin contributed to spreading the belief that Chasselas was introduced into Western Europe from Egypt or Turkey during ancient times. However, DNA analysis places Chasselas perfectly into the range of genetic variability of typical Alpine cultivars, far distant from the Eastern Mediterranean germplasm (Vouillamoz and Arnold 2009). The number of bud sports selected from Chasselas, Pinot, and Traminer is particularly high compared with other varieties of comparable diffusion, which lends support to their long history of vegetative propagation. The accumulation of somatic mutations has occurred to such an extent that some bud sports have been mistakenly considered as distinct varieties (e.g. Chasselas Doré, Chasselas Musqué, Chasselas Sans Pepins, Pinot Blanc, Pinot Menieur, Gewürztraminer, Traminer Rot). Clairette is another ancient variety that is documented in French Mediterranean coast of Languedoc since the twelfth century. The rounded seeds of Clairette suggest ties with allochthonous wild forms (Terral et al. 2010).

Alleged descendents of antique and extinct varieties were investigated using historical records as a guide. Raetica was a white grape variety mentioned by Roman authors since Pliny the Elder (23–79 AD) and widely cultivated at that time on the northern and southern edges of the Central Alps. Moving from the assumption that Raetica might have persisted throughout centuries into the etymologically similar modern variety Rezé—cultivated in the same geographical area since the fourteenth century—Vouillamoz et al. (2007) discovered a large kinship group of Alpine varieties as putative descendents of Raetica. Another valuable source of information for reconstructing the history of cultivation lies in the recovery of DNA from herbarium specimens of once cultivated varieties. A 90-year-old herbarium specimen conserved in the Natural History Museum in Split, Croatia, with the name of Tribidrag, whose cultivation has been mentioned in Dalmatian coast since the fifteenth century, turned out to match the DNA profiles of Crljenak Kaštelanski, Primitivo, and Zinfandel—the modern names by which this variety is now cultivated in Croatia, Southern Italy, and California, respectively (Malenica et al. 2011). More ancient remnants failed to provide direct genealogical links to varieties in existence today. Grape pips preserved by waterlogging or charring at archaeological sites of the Celtic, Greek, and Roman times may shed more light on historical diffusion of grapevines (Cappellini et al. 2010). DNA of seeds excavated from a Roman site (Second—fourth century AD) in the Hungarian plains provided incomplete fingerprints due to poor preservation, but data were sufficient to show ties with present-day varieties grown in Italy and Croatia (Manen et al. 2003). More ancient pips (fifth century BC) from the French Mediterranean coast provided molecular data that linked them to Austrian/German and French modern cultivars. The exact reconstruction of genealogical relationships in the extant population of grapevine is complicated by the coexistence of centuries-old grapevines and cultivars generated by more recent crosses between parents that are not coeval, and by the geographical vicinity of genetically unrelated accessions due to historical dispersal. These peculiarities make time-scaled phylogeography and population structure rather intractable for the inference of simple pedigree relationships.

27.6 Selective Sweeps

The small changes in grapevine biology imposed by domestication resulted in a limited number of signatures of selection at a genome-wide level. A locus under selection was identified on chromosome 17 in which haplotype diversity in cultivated varieties is reduced in comparison to haplotype diversity in *V. vinifera* subsp. *sylvestris* (Myles et al. 2011). The selective sweep associated with this locus extends over 5 million nucleotides, with long-range LD preventing the identification of the gene under selection. A signature of selection has not been detected where it was more likely expected—e.g. at the flower sex locus—which was previously located by bi-parental mapping on the upper arm of chromosome 2. This incongruity at the population level with the expected effect of a Mendelian and crucial domestication trait is accompanied by bizarre observations in progeny of self-pollinated hermaphroditic varieties, which always include a low percentage of female grapevines (Snyder and Harmon 1939) that persist after successive generations of selfing (Bronner and Oliveira 1990). All of this casts some uncertainty on the genetic control of flower sex. To explain the opposite phenomenon—the rare but recurrent appearance of seeded fruit set by staminate flowered vines—Barrett (1966) argued that complexes of independently segregating factors may affect sex expression through the interaction with sex genes on chromosome 2. Selective sweeps have also occurred past the event of domestication. Among cultivated varieties, haplotype diversity at the *MybA* color locus is reduced in white-skinned cultivars relative to red cultivars, confirming that most white-skinned varieties have originated by descent from a common ancestor in which the mutations at the *MybA* gene family occurred (Fournier-Level et al. 2010). Positive selection for the *MybA* white haplotype caused LD to extend far from *MybA* into a large part of the lower arm of chromosome 2, asymmetrically from downstream of the genetic centromere – near the location of *MybA*—to the telomere (Myles et al. 2011), in a fairly similar way as it occurred at the endosperm *y1* color locus in corn (Palaisa et al. 2004). Grapevine chromosome 2 has been one of the main targets of human selective pressure, with an opposite impact on the level of genetic diversity in the two chromosomal arms: LD has decayed rapidly around the sex locus on the upper arm, but long-range LD around the white-skin *Myb* mutations still persists downstream of the genetic centromere.

27.7 The Effect of Modern Breeding

Modern breeding has imparted a minimal change to global grapevine diversity, with two notable exceptions in the sectors of table grapes and disease resistant varieties. The pure *V. vinifera* varieties generated by deliberate crosses in the past 150 years have been planted on limited acreage for wine production (Owens 2008). A few novel varieties with outstanding wine quality were able to compete for cultivation with more ancient selections. For instance, Müller Thurgau (Riesling × Madeleine Royale) is appreciated in cool climates of Central Europe and has become the second most planted variety in Germany. Other varieties have become notorious for

the improvement of particular traits, e.g. the red-fleshed Alicante Bouschet for wine blends with higher anthocyanin content and deeper red color. The release of new table grape varieties has been more successful, being promptly adopted by a more dynamic market that was previously limited to a number of ancient varieties with a narrow genetic basis. Recent breeding in table grapes has promoted the expansion rather than the restriction of genetic diversity in the cultivated compartment. On the contrary, breeding for disease resistance has imposed bottlenecks due to founder effects. Intense and prolonged breeding has been conducted since the dawn of the nineteenth century to combat fungal pathogens that became dispersed across traditional areas of cultivation since that time. A small number of highly resistant North American grapevines that were also naturally free from unfavorable alleles for fruit quality were selected for initiating the hybridisation with sensitive *V. vinifera*, and entered the process of backcrossing (Munson 1909). European breeders established national programs of grapevine improvement upon those precious founders and their descendents. A significant bottleneck was imposed by the extensive exploitation of a few downy mildew resistant lineages, which turned out to each depend on distinct resistant haplotypes at a single locus called *Rpv3*, coinciding with a cluster of receptor-like resistance genes on chromosome 18 (Di Gaspero et al. 2012). Despite the fact that resistance has been introgressed into many different backgrounds of *V. vinifera*, preserving most of the original genome-wide diversity in the pool of newly created resistant varieties, this process entailed a significant selective sweep around *Rpv3*.

27.8 Genetic Diversity in Traditional Cultivars

27.8.1 Global Patterns of Genetic Structure

The early spread of domesticated forms from the Southern Caucasus occurred predominantly through seeds. Trade of grapevines via the Mediterranean Sea had a significant impact on the distribution of native communities, creating centres of accumulation of allochthonous diversity of domesticated grapevines in regions where autochthonous forms of wild diversity were part of the native flora. The ease with which grapevines hybridized in promiscuous vineyards and with nearby wild forms opened the door to gene exchange and homogenization of gene pools. When grapevine cultivars from European national germplasm collections are catalogued by the prevalent country/region of present cultivation and subjected to DNA genotyping, admixture amongst varieties of presumed different origin is extremely common (Cipriani et al. 2010; Laucou et al. 2011). Table grapes tend to separate more clearly from wine grapes as a result of their different origin and history of dispersal. The prohibition of wine consumption in Islamic countries on the southern and eastern shores of the Mediterranean since 600 AD promoted the replacement of wine grape cultivars with table grapes of the *proles orientalis* in the Middle East and North Africa, partially in the Balkans, and to a more limited extent in the Iberian peninsula.

Chloroplasts are maternally inherited in grapevine and cpDNA provides information on past changes in species distribution that remains unaffected by subsequent pollen flow. Eight different chlorotypes have been identified by chloroplast microsatellite markers, of which only four have frequencies greater than 5%, and one of them has an intermediate relationship with all others, suggesting that it might represent the ancestral chlorotype (Arroyo-García et al. 2006). All eight chlorotypes present in cultivated varieties are also found in *V. vinifera* subsp. *sylvestris* collected in Southwestern Asia, but only a variable subset of these haplotypes are found in fragmented wild populations in Europe and North Africa. Patterns of genetic structure suggest that the wild *V. vinifera* survived Quaternary glaciations in geographically isolated populations corresponding to four main refugia in the Caucasus, the Iberian and Italian peninsulas, and Sardinia. In Western Europe, the extant German population of wild grapevines in the Ketsch Island on the Rhein River shows strong links with populations in the Italian peninsula, while the Austrian population in Marchegg, farther east along the Danube River, shows admixture between southern and eastern populations. European wild populations of grapevines located on the northern side of the east-west mountain ranges were influenced by northward colonizations from the Italian peninsula via the Alps and westward colonization from refugia in the Balkan and Caucasus areas.

An east-west geographical gradient of genetic diversity among wild and cultivated grapevines has been revealed by nuclear DNA (Myles et al. 2011). The domesticated population is genetically closer to wild forms from the Near East than to western wild forms. However, western cultivars are relatively closer to western wild forms than any other cultivars. On a continental scale, admixture models indicate that western cultivars are a mixture of eastern cultivars and western wild forms, opposing the hypothesis that western wild populations are similar to western varieties because of gene introgression from the cultivated compartment. Thus, the most likely model is consistent with a single major event of domestication and introgression from local *sylvestris* as eastern cultivars moved westward into Europe and North Africa. Domestication imposed a weak bottleneck. The reduction of haplotype diversity in the domesticated population is statistically significant, but relatively weak at a genome-wide scale. This is consistent with the common beliefs that the cultivars in use today are only a few generations removed from their wild progenitors, and asexual reproduction was the predominant mode of propagation during the last millennium.

27.8.2 *The Classification of Negrul into Proles*

Negrul (1946) recognized three major groups of cultivars or *proles* based on ecological and morphological features, then referred to as *pontica*, *orientalis*, and *occidentalis* depending on their prevalent geographical distribution. *Proles pontica* comprises the oldest cultivars, which originated close to the centre of domestication and spread southward into the Middle East, westward into the Balkans, and farther west into Europe without being significantly contaminated by gene introgression from local forms. Old eastern varieties such as Rkatziteli, Saperavi, Odjalesci,

Mtsvane, Furmint, and Harslevelu belong to the *proles pontica*. Some authors have also placed in the *proles pontica* some cultivars (e.g. Clairette and Vermentino) that are currently present only in Western Europe (Mullins et al. 1992). *Proles orientalis* includes cultivars that also originated near the centre of domestication, and are presently cultivated predominantly around the southern shores of the Caspian Sea and in Central Asia. Morphological features of this group are so distinctive that Negru argued that they were domesticated from a different wild form of *V. vinifera* subsp. *sylvestris aberrans*. *Proles orientalis* includes wine grape cultivars (*Proles orientalis subproles caspica*) that were popular in the Near East before the advent of Islam, such as Arenii Tchernii, Baian Shirei, Terbash, and a mixture of wine and table grape cultivars (*proles orientalis subproles antasiatica*) such as Katta Kurgan, Nimrang, Ararati, Sultanina, and the founders of the Muscat family. *Proles occidentalis* comprises wine grapes that arose in Western Europe after the introduction of the eastern founders of the *proles pontica* and/or *orientalis subproles caspica*, which acquired adaptations to local environments by hybridization with indigenous wild forms. A more detailed classification identified taxa at lower hierarchical levels within the three *proles*, with significant diversification within the *proles orientalis* and *pontica* (Troshin et al. 1990). The first global analysis of genetic structure was based on microsatellite DNA variation among ~200 representative cultivars, which revealed four main clusters (Aradhya et al. 2003). One cluster was distinguished by the presence of seedless table grapes of the *proles orientalis subproles antasiatica*, which grouped with a few cultivars grown today along the Mediterranean shores (e.g. Sultana, Vermentino). A closely related cluster consisted of a mixture of table grapes (e.g. Black Corinth, Chaouch Blanc), wine grapes from the Greek Mediterranean shores and the Balkans (e.g. Kadarka, Limnio, Primitivo, Vranac, Xynomavro), and some notable exceptions from alleged western cultivars, such as Barbera, Carignane, Dolcetto, Folle Blanche, Ohanes, Rondinella, and Viogner. The first two clusters mainly include descendents that are presumed to be not far removed from *orientalis* and *pontica* founders. The third cluster has a heterogeneous composition in terms of the country of present cultivation, including Greek and Balkan cultivars (e.g. Harslevelu, Liatiko, Mandilaria, Mavrodaphne, Plavac Mali) and a few varieties that today are restricted to Western Europe, such as Aramon, French Colombard, Mauzac Blanc, Palomino, and Treixadura. The last cluster is more homogeneous, and includes cultivars that are argued to have received significant gene influx from western *sylvestris* (e.g. Marsanne, Merlot, Mondeuse Noir, Pinot, Sauvignon, Traminer, and their known relatives). However, the overall picture showed a low level of differentiation between groups, as well as a high level of genetic diversity (~85%) within all groups.

27.8.3 Kinship and Founders of Major Descent Groups

The prevailing conventional wisdom is that the thousands of modern cultivars encompass a tremendous proportion of ancestral diversity—an idealistic scenario functional to wine magazines for celebrating natural diversity and the historical and cultural heritage of local varieties. Analyses of genetic diversity and kinship have challenged this

customary view and clarified that diversity is contained within a network of family relationships linking most varieties one to another. It is a matter of fact that considerable phenotypic diversity does exist, but this trait variation has mostly stemmed from novel allelic combinations and recombination in heterozygous individuals within a limited number of kinship groups (Myles et al. 2011).

The number of cultivars with first-degree relationships is particularly high among table grapes, which have a narrow genetic basis. Muscat of Alexandria, which descends from Muscat Blanc à petits grains the earliest ancestor of the Muscat lineage yet discovered (Cipriani et al. 2010), and Sultanina are present in the pedigrees of many varieties. A large kinship group links by descent the table grapes Aswad, Asma, Kishmish Chernyi, Dzhandzhal Kara, Red Roumi, and Olivette Vendemien—said to be of eastern origin—and includes also wine grapes such as Mandilaria, Harslevelu, Ohanes, and Gouveio, cultivated today in Greece, Hungary, Spain, and Portugal, respectively (Myles et al. 2011). Among wine grapes, the most prolific founders of kinship groups are Traminer, Pinot, Gouais Blanc, Schiava Grossa, and Chasselas—the most ancient varieties in Western Europe.

Rare examples also exist in which a family of varieties does not match a descent group (Lacombe et al. 2007). This is the case for the Malvasia varieties, a group of old grapevines historically grown in disjuncted coastal regions across the Mediterranean Sea and in Madeira and Canary Islands off the Atlantic coast near hubs where merchants traded wines since the Middle Ages. These are named after Monemvasia or Malevizi—Greek or Cretan toponyms—transliterated into Malvasia (Italian, Spanish, and Portuguese), Malvazija (Croatian), Malvoisie (French), Malmsey (English), and frequently accompanied by a local name of the region where the variety has acquired popularity (e.g. Malvasia di Candia—the Venetian name for Crete—and Malvoisie de Madeira). However, several locally-grown varieties named as Malvasias designated by toponyms or adjectives specifying their peculiar characteristics are just homonymous of more widely known varieties (e.g. Malvasia de Manresa = Garganega, Malvoisie du Douro = Vermentino, Malvoisie Vert Petite = Riesling Italico, Malvoisie Weiss = Rezzè, Malvasia Verde = Furmint, Malvasia Lunga = Trebbiano Toscano), or offspring of geographically disjuncted parents (e.g. Malvasia del Lazio = Muscat of Alexandria × Schiava Grossa). With the exception of a small subgroup of Malvasias that share ancestry and Muscat flavors, the term Malvasia is not indicative of the genetic background of the variety. This confusion was generated during ancient times by the abuse of the designation of Malvasia for labelling exotic wines of superior quality, irrespective of their origin. The nature of the corresponding varieties remained obscure until the advent of DNA genotyping.

27.8.4 Heterozygosity

Self-compatibility and cleistogamy are common characters in hermaphroditic cultivars of *V. vinifera*, which have a predominantly autogamous mating system (Chkhartishvili et al. 2006). Contrary to expectations, DNA genotyping proved that

very few popular varieties have originated by selfing (Cipriani et al. 2010). This is particularly surprising, as most varieties were originally chance seedlings that arose in the vineyards (Bowers et al. 1999). Self- or cross-pollination are equally efficient in setting seeds, but seed viability and germination rates are reduced upon self-pollination (Sabir 2011). Literature support for this issue is scarce, but grapevine breeders empirically experience that progeny of selfed plants grow weakly, and juvenile plants are more likely to succumb to severe environmental conditions than those resulting from outcrossed seeds. Inbreeding depression is usually severe, with a rapid drop in vigor and fertility within a few generations of forced autogamy (Bronner and Oliveira 1990). The decline in fitness of seedlings that originated from selfing has probably disfavored their retention in cultivation due to juvenile mortality, low vigour and low fertility of the survivors. Grapevine has therefore maintained high levels of heterozygosity during domestication and selection of new cultivars, which is evident in genome-wide DNA scans and in the extent of phenotypic variation in progeny of self-pollinated varieties (Snyder and Harmon 1939; Myles et al. 2010). The heterozygous state is also associated with a genetic load of recessive deleterious alleles that becomes apparent upon self-pollination. The intolerance to inbreeding supports the claim that heterosis is responsible for the large success of spontaneous crosses between genetically dissimilar varieties that generated sibling groups including the world's finest varieties (Bowers et al. 1999).

Near-homozygous grapevines were produced experimentally from Pinot Noir through successive generations rounds of self-pollination intercalated by the elimination of the weakest seedlings in each generation (Bronner and Oliveira 1990). This process was intended to remove part of the genetic load and to counteract the rise in mortality and sterility that accompanied the increase of homozygosity. The final level of homozygosity after reiteration in selfing ranged between 75 and 97 % among different lines, and was the highest in PN40024, the individual chosen for whole-genome sequencing (Jaillon et al. 2007). The recorded pedigree of PN40024 indicated that near homozygosity was achieved after nine generations of selfing, but DNA analysis revealed an accidental outcrossing with Helfensteiner—itself an offspring of Pinot Noir and Schiava Grossa—during one of the early generations. We have resequenced the trio Pinot Noir, Schiava Grossa, and PN40024 (authors, unpublished data) showing that the genomes of Pinot Noir and Schiava Grossa have very small homozygous blocks and the complementary heterozygous regions are prevalent (86 and 82 % of 50-kb windows across the genome, respectively), and this heterozygosity became reduced to ~5 % in PN40024. Part of the residual heterozygosity was concentrated around the flower sex locus, which was maintained in the heterozygous state by the deliberate selection of hermaphroditic individuals before every round of selfing.

27.8.5 DNA Sequence Diversity

The estimation of genetic diversity in the most comprehensive sample of grape germplasm is based on 3,727 accessions of *V. vinifera* subsp. *sativa* and 80 individuals of *sylvestris* analysed at 20 microsatellite loci (Laucou et al. 2011). As many

as 2,227 *sativa* accessions (60%) corresponded to unique cultivars, the remainder were redundant due to clonal variation or synonymy. The mean genetic diversity index was 0.769 ± 0.121 in *sativa* accessions. The cultivated portion of grapevine germplasm is as diverse as poplar, roses, and corn, and much more diverse than tomatoes and wheat. The analysis of nine distantly related grapevine cultivars and two *sylvestris* accessions revealed high levels of nucleotide diversity ($\pi = 0.0051$) in a sample of 230 genes totalling ~ 2 million nucleotides—making the level of DNA polymorphism in grapevines comparable to that of outcrossing crops and much higher than that of autogamous crops (Lijavetzky et al. 2007). Whole-genome sequencing of the heterozygous Pinot Noir provided the first global inventory of nucleotide diversity between two grape haplotypes, which were estimated to differ by an average of four single nucleotide polymorphisms (SNP) per kilobase (Velasco et al. 2007). In addition to the private SNPs found in Pinot Noir, genome-wide SNP discovery was initiated by resequencing another 15 varieties using sequencing-by-synthesis technology (Myles et al. 2010). A 71K SNP set was identified from a $\sim 2.3\%$ fraction of the grapevine genome (Myles et al. 2010). All of these data confirm that grapevine has a highly heterozygous genome. Self-compatibility and cleistogamy that appeared during domestication had no time or way to increase the level of genome-wide homozygosity in cultivated grapevines compared to their obligatory allogamous ancestors, either because of lower fitness and breeding value of selfed seedlings or because of the practice of perpetuating favorable genotypes through vegetative propagation, which prevented reassortment past a few generations of mating.

27.8.6 Haplotype Diversity

Genotyping of a thousand grapevines at 5,387 SNPs has provided the most precise estimation of genome-wide haplotype diversity in grapevine (Myles et al. 2011). More than a thousand of these SNPs were physically located at an inter-SNP distance of < 1 kb, which provided high resolution power. Haplotype diversity in genomic windows of > 5 SNPs was significantly higher in *sylvestris* than *sativa*, but in absolute terms the observed difference was negligible. The same holds true within the cultivated compartment (in windows > 10 SNPs) between cultivars of the *proles pontica* and *orientalis*, originating close to the centre of domestication, and cultivars of the *proles occidentalis* typical of Western Europe. Regardless of this classification, haplotype diversity is always > 0.9 in windows of > 10 SNPs. The above mentioned analysis of Lijavetzky and coworkers (2007) of a sample of 230 genes in 11 diploid grapevines detected an average of 6.6 haplotypes per gene—over a maximum expected sampling of 22 homologous chromosomes—with the most common case being the presence of a major haplotype with an average frequency of 0.49 and the first three haplotypes accounting for an average cumulative frequency of 0.83 (Lijavetzky et al. 2007). More precise estimates of haplotype diversity are yet to come from the complete

resequencing of a substantial number of grapevine accessions. Should resequencing be extended to a significant proportion of the diversity present in *sativa* and *sylvestris*, signatures of the domestication syndrome may also become more easily detectable.

27.8.7 *Patterns of Linkage Disequilibrium*

The decay of linkage disequilibrium (LD) occurs rapidly; r^2 between SNPs drops down to background level within 1 kb without a significant difference between wild and cultivated forms (Myles et al. 2011). In contrast, microsatellite loci indicate significant LD at a centiMorgan (cM) scale. In a diversity panel of 141 cultivated varieties, r^2 declined to 0.1 within ~ 10 cM, which should correspond to 1.3–2.2 million nucleotides (Barnaud et al. 2006), and in wild forms within 2.7 cM (Barnaud et al. 2010). This apparent inconsistency is probably due to the combination of three factors: the way LD is measured, the different causes of LD decay that SNP and microsatellite markers are able to capture, and the different set of genotypes used. In grapevine, limited recombination past the event of domestication and family relationships should have favored the persistence of large haplotype blocks among cultivars, which argue against rapid LD decay. Vegetative propagation of interesting genotypes has immortalized existing LD blocks, but also increased the role played by mutation in disrupting LD. LD measured in terms of D' reflects only recombinational history, while r^2 summarizes both recombinational and mutational history. The decay rate is faster for r^2 than D' and, in contrast with D' , r^2 also declines within haplotype blocks shared by descent, provided that mutations have accumulated past the event of radiation. High density SNPs are better suited than microsatellites to capture mutations that contributed to LD decay, as estimated by r^2 .

27.8.8 *Structural Diversity*

Homologous chromosomes are highly dissimilar in extensive DNA stretches flanked by orthologous regions. This structural variation has been detected in the sequence of the heterozygous variety Pinot Noir (Velasco et al. 2007), either as long non-colinear DNA stretches located in orthologous positions in the two chromosomes—i.e. two completely different stretches of DNA sequence present in the same position on the two haplotypes, amounting to ~ 65 million nucleotides—or as presence/absence variation—i.e. a sequence that is found in one haplotype and is absent from the other, amounting to ~ 49 million nucleotides. A large fraction of the structural differences appears to be directly or indirectly mediated by repetitive/transposable element (TE) activity. Repeats and transposable elements represent as much as 41.4 % of the grapevine genome, and introns are particularly enriched in long interspersed element (LINE) retrotransposons, most of which are still transcriptionally active in grapevine (Jaillon et al. 2007). An uncharacterized part of structural variation involves DNA

stretches much longer than TEs, which in some instances are hundreds of kilobases long and contain dispensable genes and/or redundant gene copies. The biological significance of structural diversity is still largely unexplored and only the results from individual genome resequencing will shed light on the processes that have generated this variation and its relevance for phenotypic diversity.

Acknowledgment The authors thank Courtney Coleman for proofreading the manuscript.

References

- Aradhya MK, Dangi GS, Prins BH et al (2003) Genetic structure and differentiation in cultivated grape, *Vitis vinifera* L. *Genet Res* 81:179–192
- Arrigo N, Arnold C (2007) Naturalised *Vitis* rootstocks in Europe and consequences to native wild grapevine. *PLoS One* 2(6):e521
- Arroyo-García R, Ruiz-García L, Bolling L et al (2006) Multiple origins of cultivated grapevine (*Vitis vinifera* L. ssp. *sativa*) based on chloroplast DNA polymorphisms. *Mol Ecol* 15:3707–3714
- Barnaud A, Lacombe T, Doligez A (2006) Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L. *Theor Appl Genet* 112:708–716
- Barnaud A, Laucou V, This P et al (2010) Linkage disequilibrium in wild French grapevine, *Vitis vinifera* L. subsp. *silvestris*. *Heredity* 104:431–437
- Barrett HC (1966) Sex determination in a progeny of a self pollinated staminate clone of *Vitis*. *Proc Am Soc Hortic Sci* 88:338–340
- Barth S, Forneck A, Verzeletti F (2009) Genotypes and phenotypes of an *ex situ* *Vitis vinifera* ssp. *silvestris* (Gmel.) Beger germplasm collection from the Upper Rhine Valley. *Genet Resour Crop Evol* 56:1171–1181
- Bouquet A (1980) *Vitis* × *Muscadinia* hybridization: a new way in grape breeding for disease resistance in France. *Proc. 11h 6 Intern. Symp. Grape Breed.*, 15–18 June 1980, Davis, USA, 42–61
- Boursiquot JM, Lacombe T, Bowers JE, Meredith CP (2004) Le Gouais, un cépage clé du patrimoine viticole européen. *Bulletin de l'OIV* 77 (875–876):5–19
- Bowers J, Boursiquot J-M, This P et al (1999) Historical genetics: the parentage of Chardonnay, Gamay, and other wine grapes of Northeastern France. *Science* 285:1562–1565
- Brantjes NBM (1978) Pollinator attraction of *Vitis vinifera* ssp. *silvestris*. *Vitis* 17:229–233
- Bronner A, Oliveira J (1990) Creation and study of the Pinot noir variety lineage. *Vitis Special Issue*: 69–80
- Caporali E, Spada A, Marziani G et al (2003) The arrest of development of abortive reproductive organs in the unisexual flower of *Vitis vinifera* ssp. *silvestris*. *Sex Plant Reprod* 15:291–300
- Cappellini E, Gilbert MT, Geuna F et al (2010) A multidisciplinary study of archaeological grape seeds. *Naturwissenschaften* 97:205–217
- Chkhartishvili N, Vashakidze L, Gurasashvili V, Maghradze D (2006) Type of pollination and indices of fruit set of some Georgian grapevine varieties. *Vitis* 45:153–156
- Cipriani G, Spadotto A, Jurman I et al (2010) The SSR-based molecular profile of 1005 grapevine (*Vitis vinifera* L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theor Appl Genet* 121:1569–1585
- Detjen LR (1919) The limits in hybridisation of *Vitis rotundifolia* with related species and genera. *NC Agri Expt Sta Bull* 12
- Di Gaspero G, Copetti D, Coleman C et al (2012) Selective sweep at the *Rpv3* locus during grapevine breeding for downy mildew resistance. *Theor Appl Genet*. doi:10.1007/s00122-011-1703-8

- Di Vecchi-Staraz M, Laucou V, Bruno G et al (2009) Low level of pollen-mediated gene flow from cultivated to wild grapevine: consequences for the evolution of the endangered subspecies *Vitis vinifera* L. subsp. *sylvestris*. *J Hered* 100:66–75
- El Oualkadi A, Ater M, Messaoudi Z et al (2011) Genetic diversity of Moroccan grape accessions conserved *ex situ* compared to Maghreb and European gene pools. *Tree Genet Genome*. doi:10.1007/s11295-011-0413-3
- Fairon-Demaret M, Smith T (2002) Fruit and seeds from the Tienen formation at Doornal Palaeocene-Eocene transition in Eastern Belgium. *Rev Paleobot Palynol* 122:47–62
- Fournier-Level A, Lacombe T, Le CL (2010) Evolution of the *VvMybA* gene family, the major determinant of berry colour in cultivated grapevine (*Vitis vinifera* L.). *Heredity* 104:351–362
- Gallardo A, Ocete R, Ángeles LM (2009) Assessment of pollen dimorphism in populations of *Vitis vinifera* subsp. *sylvestris* (Gmelin) Hegi in Spain. *Vitis* 48:59–62
- Gong F, Karsai I, Liu Y-S (2010) *Vitis* seeds (Vitaceae) from the late Neogene Gray Fossil Site, northeastern Tennessee, U.S.A. *Rev Paleobot Palynol* 162:71–83
- Jackson RS (2000) Wine science: principles, practise, perception. Academic Press, San Diego 648 pp
- Jaillon O, Aury JM, Noel B et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449:463–467
- Jansen RK, Kaittani C, Saski C et al (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. *BMC Evol Biol* 6:32
- Jelenković G, Olmo HP (1968) Cytogenetics of *Vitis*. III Partially fertile F₁ hybrids between *V. vinifera* L. × *V. rotundifolia* Michx. *Vitis* 7:281–293
- Lacombe T, Boursiquot J-M, Laucou V et al (2007) Relationships and genetic diversity within the accessions related to Malvasia held in the Domaine de Vassal grape germplasm repository. *Am J Enol Vitic* 58:124–131
- Laucou V, Lacombe T, Dechesne F et al (2011) High throughput analysis of grape genetic diversity as a tool for germplasm collection management. *Theor Appl Genet* 122:1233–1245
- Levadoux L, Boubals D, Rives M (1962) Le genre *Vitis* et ses especes. *Annales Amelioration des Plantes* 12:19–44
- Lijavetzky D, Cabezas JA, Ibáez A et al (2007) High throughput SNP discovery and genotyping in grapevine (*Vitis vinifera* L.) by combining a re-sequencing approach and SNPlex technology. *BMC Genomics* 8:424
- Malenica N, Šimon S, Besendorfer V et al (2011) Whole genome amplification and microsatellite genotyping of herbarium DNA revealed the identity of an ancient grapevine cultivar. *Naturwissenschaften* 98:763–772
- Manen JF, Bouby L, Dalnoki O et al (2003) Microsatellites from archaeological *Vitis vinifera* seeds allow a tentative assignment of the geographical origin of ancient cultivars. *J Archaeol Sci* 30:721–729
- Miller NF (2008) Sweeter than wine? The use of the grape in early Western Asia. *Antiquity* 82:937–946
- Moore MJ, Soltis PS, Bell CD et al (2010) Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc Natl Acad Sci USA* 107:4623–4628
- Mullins M, Bouquet A, Williams L (1992). *Biology of the grapevine*. Cambridge University Press, NY
- Munson TV (1909) *Foundations of American grape culture*. T.V. Munson & Son, Denison
- Myles S, Chia JM, Hurwitz B et al (2010) Rapid genomic characterization of the genus *Vitis*. *PLoS One* 5:e8219
- Myles S, Boyko AR, Owens CL et al (2011) Genetic structure and domestication history of the grape. *Proc Natl Acad Sci USA* 108:3530–3535
- Negi SS, Olmo HP (1966) Sex conversion in a male *Vitis vinifera* L. by a kinen. *Science* 152:1624–1625
- Negi SS, Olmo HP (1970) Studies on sex conversion in male *Vitis vinifera* L. *sylvestris*. *Vitis* 9:89–96

- Negi SS, Olmo HP (1971) Conversion and determination of sex in *Vitis vinifera* L. *sylvestris*. *Vitis* 9:265–279
- Negrul AM (1946) Origin and classification of cultivated grape. In: Baranov A, Kai YF, Lazarevski MA, Palibin TV, Prosmoserdov NN (eds) The Ampelography of the USSR. Volume 1. Pischepromizdat, Pischepromizdat, Moscow, 159–216
- Ocete R, Arnold C, Failla O et al (2011) Considerations on the European Wild grapevine (*Vitis vinifera* L. ssp. *sylvestris* (Gmelin) Hegi) and Phylloxera infestation. *Vitis* 50:97–98
- Olmo HP (1986) The potential role of *V. vinifera* × *rotundifolia* hybrids in grape variety improvement. *Experientia* 42:921–926
- Olmo HP (1996) The origin and domestication of the *vinifera* grape. In: McGovern P, Fleming SJ, Katz SH (eds) The origin and ancient history of wine. Gordon and Breach Publishers, pp 31–43
- Owens CL (2008) Grapes. In: Hancock JF (ed) Breeding temperate fruit crops: germplasm to genomics. Kluwer Academic Publishers
- Palaisa K, Morgante M, Tingey S, Rafalski A (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize *Y1* gene are indicative of an asymmetric selective sweep. *Proc Natl Acad Sci USA* 101:9885–9890
- Patel GI, Olmo HP (1955) Cytogenetics of *Vitis*: I. The hybrid *V. vinifera* × *V. rotundifolia*. *Am J Bot* 42:141–159
- Pauquet J, Bouquet A, This P, Adam-Blondon A-F (2001) Establishment of a local map of AFLP markers around the powdery mildew resistance gene *Run1* in grapevine and assessment of their usefulness for marker assisted selection. *Theor Appl Genet* 103:1201–1210
- Péros J-P, Berger G, Portemont A et al (2010) Genetic variation and biogeography of the disjunct *Vitis* subg. *Vitis* (Vitaceae). *J Biogeography* 38:471–486
- Perret M, Arnold C, Gobat J-M, Küpfer P (2000) Relationships and genetic diversity of wild and cultivated grapevines (*Vitis vinifera* L.) in central Europe based on microsatellite markers. *Acta Hort* 528:155–159
- Regner F, Stadlhuber A, Eisenheld C, Kaserer H (2000) Considerations about the evolution of grapevine and the role of Traminer. *Acta Hort* 528:177–181
- Riaz S, Tenscher AC, Smith BP et al (2008) Use of SSR markers to assess identity, pedigree, and diversity of cultivated muscadine grapes. *J Amer Soc Hort Sci* 133:559–568
- Sabir A (2011) Influences of self-and cross-pollinations on berry set, seed characteristics and germination progress of grape (*Vitis vinifera* cv. Italia). *Int J Agric Biol* 13:591–594
- Soltis DE, Smith SA, Cellinese N et al (2011) Angiosperm phylogeny: 17 genes, 640 taxa. *Am J Bot* 98:704–730
- Snyder E, Harmon FN (1939) Grape progenies of self-pollinated *vinifera* varieties. *Proc Am Soc Hort Sci* 37:625–626
- Terral JF, Tabard E, Bouby L et al (2010) Evolution and history of grapevine (*Vitis vinifera*) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Ann Bot* 105:443–455
- This P, Lacombe T, Thomas MR (2006) Historical origins and genetic diversity of wine grapes. *Trends Genet* 22:511–509
- Tröndle D, Schröder S, Kassemeyer HH et al (2010) Molecular phylogeny of the genus *Vitis* (Vitaceae) based on plastid markers. *Am J Bot* 97(7):1168–1178
- Troshin LP, Nedov PN, Litvak AI, Guzun NI (1990) Improvement of *Vitis vinifera sativa* D.C. taxonomy. *Vitis Special Issue*:37–43
- Turner C (1968) A note on the occurrence of *Vitis* and other new plant records from the Pleistocene deposits at Hoxne, Suffolk. *New Phytol* 67:333–334
- Vavilov NI (1951) The origin, variation, immunity and breeding of cultivated plants. Translation from the Russian by Chester KS. *Chronica Botanica*; Stechert-Hafner, New York, p 364
- Velasco R, Zharkikh A, Troggio M et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* 2:e1326
- Vouillamoz JF, Schneider A, Grando MS (2007) Microsatellite analysis of Alpine grape cultivars (*Vitis vinifera* L.): alleged descendants of Pliny the Elder's Raetica are genetically related. *Genet Resour Crop Evol* 54:1095–1104

- Vouillamoz J, Arnold C (2009) Etude historico-génétique de l'origine du 'Chasselas. *Revue suisse Vitic Arboric Hortic* 41: 299–307
- Walker AR, Lee E, Bogs J et al (2007) White grapes arose through the mutation of two similar and adjacent regulatory genes. *Plant J* 49:772–785
- Wan Y, Schwaninger H, Li D et al (2008a) A review of taxonomic research on Chinese wild grapes. *Vitis* 47:81–88
- Wan Y, Schwaninger H, Simon CJ et al (2008b) The eco-geographic distribution of wild grape germplasm in China. *Vitis* 47:77–80
- Westergaard M (1958) The mechanism of sex determination in dioecious flowering plants. *Adv Genetics* 9:217–281
- Zecca G, De Mattia F, Lovicu G et al (2010) Wild grapevine: silvestris, hybrids or cultivars that escaped from vineyards? Molecular evidence in Sardinia. *Plant Biol (Stuttg)* 12:558–562
- Zecca G, Abbott JR, Sun WB, Spada A, Sala F, Grassi F (2011) The timing and the mode of evolution of wild grapes (*Vitis*). *Mol Phylogenet E* 62:736–747
- Zohary D (1996) The domestication of *Vitis vinifera* L. in the Near East. In: McGovern P, Fleming SJ, Katz SH (eds) *The origin and ancient history of wine*. Gordon and Breach Publishers, Amsterdam, pp 21–28

Index

IBL.1RS translocation, 264
454 Titanium, 534, 543, 546–548, 550–552

A

Adaptive traits, 68, 69, 71, 76
admixture values, 497
advanced-backcross QTL analysis, 501
Aegilops biuncialis, 267
Aegilops cylindrica, 267
Aegilops species, 266
AFLP, 486, 534, 535
A-genome, 338, 339
Allelic loss, 71
Allelic richness, 70, 75, 78–81
Allelic shifts, 71
Allopolyploid, 121–123
allozymes, 486
amiprophos-methyl, 331
Annona cherimola, 68
ancestor, 691
ancestral population, 489
Andean gene pool, 487
aneuploid mutants, 328
Annotation, 406, 538, 541, 543, 545, 547
Assembly, 391, 394, 396, 401, 406, 544–548
assembly methods, 325
association genetics, 501
association mapping, 97
Arabidopsis, 174, 325, 326
autogamous, 699
autogamy, 698
Autopolyploid, 121

B

BAC end sequences, 412
BAC-by-BAC, 415
backcrossing, 51
Bacterial Artificial Chromosome (BAC), 329, 392, 409

Barley, 115, 119, 123, 178, 326, 391, 426, 586–588, 590, 594, 597–601, 603, 604, 606
bean germplasm, 495
beta diversity, 79, 81
Best Linear Unbiased Prediction, 96
B-genome, 338, 339
bioenergy, 604
Biogeography, 68
bioinformatic, 406, 525, 538, 599, 600, 602
BLASTX, 538, 545, 547
BLUP, 96
bottleneck, 70, 487, 490
Brassica, 121, 123
Brachypodium, 394, 411
bread wheat, 326, 328
breeding, 173, 693
bulked segregant analysis, 98
burden of proof, 56

C

(cp)SSR data, 487
C. cardunculus, 533, 534, 536, 537, 540–543, 545–553
chromosome approach, 327
candidate gene, 98, 489
cardoon, 560
cDNA libraries, 500
CEL I, 350
Celery Juice Extract, 350
cell cycle synchronization, 327
cell sorter, 323, 329
center of domestication, 491
Centres of domestication, 73
centromeres, 327
Cherimoya, 68, 73, 75–80, 83–86
Cherimoya genetic resources, 78
Chinese Spring, 329
Chloroplast alleles, 72
chromosome, 394–396, 406

- chromosome 3B, 410
 chromosome approach, 326–329
 chromosome banding technique, 256
 chromosome discrimination, 327
 chromosome genomics, 328, 329
 chromosome isolation, 327
 chromosome size, 329
 Chromosome sorting, 335, 344
 chromosome suspension, 336, 344
 Chromosome-mapped markers, 76
 Chromosomes, 327
 chromosomes in suspension, 329, 338
 CIAT, 501
 Circular neighbourhood, 72, 75, 78, 79, 81, 82, 84
 Circular neighbourhood approach, 72
 cleistogamy, 697
 Climate change, 68, 75–77, 79, 83–85
 Climate change impact, 75–77, 83, 84
 Climate change threats, 75
 climate data, 73, 78
 climate models, 79, 85, 86
 clustering methods, 497
 CNV, 326, 327
 co-retention, 293
 collinear, 411
 colonization, 687
 common bean, 484, 485, 497
 Comparative genomic, 308
 comparative genomics approach, 497
 comparative legume genomics, 497
 Conifers, 326
 consensus map, 490
 conservation genetics, 68, 485
 Conservation genomics, 68
 Conservation strategies, 69, 72
 conservation unit, 72
 contig, 394, 396, 397, 399, 400, 415, 534, 538, 539, 544, 545, 547, 548, 550, 552
 contig CL4773Contig1, 539
 Copy number variation, 123, 326, 344
 corn, 326
 cosmetic breeding, 61
 cpSSR markers, 495
 Creso, 344
 crop, 181, 484, 684
 Crop diversity, 72, 74
 crop varieties, 68
 Cross-pollination, 690
 cultivated (*Glycine max*) and wild soybean (*G. soja*), 466
 Cyt3, 331
 Cynara, 560
Cynara cardunculus, 533, 534, 542
 Cytogenetic stocks, 303
 cytogenetics, 320, 328
 cytoplasmic bottleneck, 495
- D**
 β -D-glucan, 261
 DAPI, 322, 336, 338
 Data integration, 406
 data warehouses, 430
Dasyphyrum villosum (L.), 331
de novo sequencing, 326, 344
 D-genome, 339, 344
 deflection, 325
 demographic models, 491
 denaturation, 329
 Denaturation of DNA, 331
 Diffusion models, 74
 disease resistance, 687
 Distinctness, 33
 ditelosomic lines, 329
 DIVA-GIS, 78
D. villosum, 343
 diversification, 74, 484, 486, 686
 diversity analyses, 485
 Diversity hotspot, 83
 Diversity studies, 502
 DNA Conservation, 577
 DNA content, 326, 329, 336, 340
 domesticated gene pools, 491
 domesticated populations, 492
 Domestication, 68–71, 73, 74, 77, 123, 125, 126, 465, 483–485, 489, 492, 502, 684
 domestication bottleneck, 490
 domestication events, 490
 domestication syndrome, 489, 490
 domestication-related traits (DRTs), 465
 dot plot, 325, 336
 Double-Haploid, 50
 double-strand breaks, 287
 duplications, 411
- E**
 Ecotilling, 353
 Effective population size, 70
 Environmental Envelope Modelling (EEM), 73
 Enzymatic mismatch cleavage, 350
 epigenetic marker, 521
 Epigenomics, 521
 Essentially Derived Variety (EDV), 52
 EST, 409, 485, 499, 500, 538, 545–547, 550, 552
 EST-SSR, 538, 539, 541

Evolution, 412, 483, 502, 587–591, 595, 596, 599–601, 606, 685
 Evolution of *P. vulgaris*, 492
 Evolutionary processes, 69, 70, 71

F

farmer selection, 494
 FCY, 327
 federated database, 430
 fertility, 686
 field, 174
 FISH, 320, 321, 329, 331
 FISH karyotype of the E genome of *Elytrigia elongata*, 272
 FISHIS, 330, 331, 337, 339, 340
 FITC, 331
 Fitness, 70, 76
 flow chamber, 325
 Flow cytogenetics, 327, 343
 flow cytogenomics, 343
 flow cytometer, 321, 329
 Flow Cytometry, 321, 330, 335
 flow karyotyping, 327
 Flow Sorting, 321, 322, 327, 340, 406
 Fluorescence *in situ* hybridization (FISH), 257
 fluorescent oligos, 340
 founder, 493, 687
 founder effect, 76, 489
 frost hardiness, 687
 functional genetic variation, 70
 Functional Genomics, 174, 357, 499, 500
 future, 70, 75–77, 79, 83–85

G

(GAA)₇, 338, 340
 GISH, 320
Genlisea, 326
 gene banks, 485
 gene density, 411
 gene flow, 70, 485, 491, 493
 Gene Function, 592, 601, 602
 gene islands, 412
 Gene Ontology, 541, 547
 gene pools, 486
 Gene silencing, 594, 602
 gene-environment interactions, 191
 genetic conformity, 56
 Genetic distinctiveness, 70
 Genetic diversity, 58, 104, 483, 490
 Genetic diversity hotspots, 73
 genetic drift, 485
 genetic engineering, 55
 genetic erosion, 68, 71, 75, 77, 83, 690

genetic gain, 28
 genetic linkage, 500
 genetic map, 498, 515, 533, 537, 542, 543, 553
 genetic markers, 97, 515
 Genetic responses, 71
 Genetic structure, 74, 81, 82
 Genome sequence, 406, 591, 597, 599, 604
 Genome sequencing, 117, 126, 127, 325
 genome structure, 427
 Genome wide association (GWAS), 114
 Genome-wide association mapping, 485
 genome-wide mapping, 76
 genome wide selection, 100
 genome-wide transcript studies, 500
 GenomeZipper, 427
 Genomic *in situ* hybridization (GISH), 257, 320
 genomic selection, 114, 127
 genomic tools, 484, 497
 Genotyping-by-sequencing, 127
 Geographic Information Systems, 68
 Geographic patterns of diversity, 73
 geographical distribution, 485
 geographical structures, 486
 Georeferenced observation points, 73
 Georeferenced plant data, 73
 Geospatial analyses, 73
 Geospatial analysis, 72
 Germplasm, 50, 69, 71, 76–78, 485, 564, 591, 592, 595, 606
 Germplasm Characterization, 355
 Germplasm collection, 68, 484, 497
 GIS, 72, 74
 Globe artichoke, 533–538, 541–546, 551, 552, 562,
 GO categorisation, 538
 GO terms, 539
 Good Phenotyping Practice (GPP), 185
 grape, 326
 Grapevine, 683
 growth habit, 489
 GWS, 101, 103, 106

H

Haplotype, 118, 121, 122, 124–126, 686
 Heterochromatin, 124
 heterosis, 698
 High-resolution physical map, 287
 high-throughput genomic technologies, 485
 high-throughput genotyping, 498
 high-throughput selection, 502
 homeologous genomes, 328, 344
 homeologous chromosomes, 326

Homologous recombination, 287
Hotspots of genetic diversity, 84
hybridization, 338, 487, 496, 690
Hybrids, 496, 565

I

γ -irradiation, 270
In situ conservation, 68, 70
In situ fluorescent hybridization, 321, 329
in situ hybridization, 330
Illumina, 533, 534, 542–544, 546–548, 550, 552
Illumina GoldenGate assay, 501
imaging spectroscopy, 184
in vitro storage, 564
inbreeding, 698
Inbreeding depression, 69
Indel, 350
intellectual property protection (IPP), 28, 50
intercrossable, 486
intergenomic translocations, 270
International Wheat Genome Sequencing Consortium (IWGSC), 328, 407
Interspecific Hybridization, 256
introgression, 491, 501, 684

K

k-mers, 419
karyotype, 422
KASPar technology, 501
kinship, 70, 697

L

Locally common alleles, 70
landraces, 68, 70, 71, 494
lignocellulosic biomass, 562
Linkage disequilibrium, 114, 118, 123, 124, 126
linkage disequilibrium analysis, 498
linkage drag, 99
linkage mapping, 97
local adaptation, 485
loss of diversity, 487
LTR retroelements, 545
lysis buffer, 331

M

map-based cloning, 597–599
mapping populations, 429, 498
marker assisted backcrossing (MABC), 98
Marker Assisted Recurrent Selection (MARS), 104

marker loss, 295
marker-assisted selection, 502
MARS, 106
MAS, 104, 106
maximum likelihood approach, 496
MDA, 343, 344
metabolomics, 527, 528
metaphase, 327, 331
metaphase chromosomes, 321
Microsatellite, 329, 340, 344
microsatellite DNA probe, 343
Microsatellite markers, 68, 78
minimal tiling path, 414
miRNAs, 548
mitotic index, 327
model plant, 603
model species, 587, 588, 591, 594–596, 599, 602, 604, 605
Molecular Assessment, 569
molecular marker, 28, 55, 344, 407, 500, 586, 591, 595–598, 603
Morphological Characterization, 71, 490, 567
multilocus molecular markers, 498
multilocus sequence data, 491
multiple domestications, 491
mutagen, 286
mutagenesis, 290
mutation breeding, 50

N

N. sylvestris, 512, 513, 515, 519, 524
N. tomentosiformis, 512, 513, 515, 519, 524
N. tabacum, 512, 513, 515, 518, 520–522, 524, 527
Neighbourhood-by-distance, 79
Next Generation Sequencing, 325, 343, 391, 392, 394, 395, 400, 406, 465, 501, 533, 542, 543, 552
nick-translation, 331
non collinear genes, 412
non-homologous end joining, 287
nozzle tip, 325
nucleic acids probes, 321
nucleotide diversity, 487, 498

O

Oryza, 325
oligonucleotides, 335
On farm conservation, 69
Optical mapping, 429
ornamental plants, 191
orthologous, 700

P

P. acutifolius, 484
P. coccineus, 484
P. lunatus, 484
P. vulgaris, 484, 486, 497
Paris japonica, 326
P. dumosus, 484
panel, 299
pangenome, 326
pedigree, 51, 691
Perennial tree crops, 73
perennials, 687
petunia, 178
phaseolins, 492
Phaseolus, 484
Phaseolus vulgaris, 483
Phaseomics, 498
phenylpropanoids, 536
photoperiod sensitivity, 489
phylogenetic, 589, 600
physical map, 286, 397, 400, 406
plagiarism, 52
plant architecture, 600, 605
Plant Breeders' Rights (PBR), 32
Plant breeding, 114, 123, 126, 496
plant evolution, 326
Plant Variety Protection (PVP), 28
Plastic responses, 76
pollination, 77, 83
polymorphism, 422
Polyploid, 406, 594
polyploidization, 596
polyploidy, 325, 326, 344
poplar, 178
population structure, 487
positional cloning, 304
Post-glacial migration routes, 73
potato, 326
Predominant derivation, 55
principal coordinate analysis, 497
Proteomics, 528
pseudogenes, 427
pseudomolecule, 429
pTa71, 331, 340
purity, 342, 344
purity checking, 340

Q

QTL analysis, 498, 500
QTL Pyramiding, 98
quantitative trait loci (QTL), 97
quantitative trait locus (QTL), 489

R

γ -rays, 290
races, 491
RAD, 533, 542–545, 551–553
Radiation, 286
radiation hybrid, 285, 286
RADseq, 543
RADSequencing, 543
random amplified polymorphic DNA, 486
rDNA, 344
Re-sampling without replacement, 73, 78, 79, 81
rearrangements, 411
Recalcitrant seed, 68
recombinant inbred population, 489
recombination, 293, 487, 700
Reduced complexity, 116, 126
Reference genome, 122, 125, 126
refugia, 70, 73
relict population, 489
remote sensing, 179
renewable energy, 537
repetitive DNA, 328, 392, 393, 400, 401
repetitive DNA probe, 344
repetitive sequences, 331
Resistance Gene Analogs, 548
resource-use efficiency, 173
restriction fragment length polymorphism, 486
retention/loss frequency, 297
retrotransposon, 412, 587, 591
reverse genetics, 307
RFLP, 586
RGAs, 548
Rice, 326, 394, 411
root tip, 327, 331
roots, 174
rRNA, 327

S

Sanger sequencing, 325, 391, 392
satellites, 327
scaffold, 396–398, 400, 415
Seed banks, 68
seed dormancy, 489
seed oil, 562
seed propagated, 564
seed protein phaseolin, 486
seed proteins, 486
seed system, 71
selection, 407, 490, 493, 502, 693
Selective sweeps, 693
Self-compatibility, 697

- Sequence capture, 116
sequence data, 487
sequencing, 485
sesquiterpene lactones, 536
SHATTERPROOF 1, 489
sheath fluid, 323
short repeats, 327
short tandem repeats, 327
shotgun sequencing, 394, 397, 400
signature of domestication, 490
Simple sequence repeat, 68
single nucleotide polymorphism (SNP), 28, 100
SNP, 325, 350, 426, 485, 538, 541–543, 545–548, 550–552
SNP Calling, 119, 545, 546, 548, 550
SNP discovery, 501
SNPs, 533, 534, 538, 542, 543, 545, 546, 550–553
somatic mutations, 692
sorghum, 411
sorting chromosomes, 323
sorting gate, 325, 341
sorting windows, 336
soybean, 497
soybean genome, 497
Spatial genetics, 68
spatial isolation, 492
Spatial principle component analysis, 83
speciation, 484
species distribution, 69
SSR, 78, 329, 533, 535, 538–541
standard operation protocols (SOP), 185
structural variation, 700
Structure analysis, 497
Survey Sequencing, 394, 426
Suspensions of Plant Chromosomes, 331
synteny, 498, 586–590, 593, 596, 597, 599, 600, 603, 685
- T**
T. aestivum, 329, 344
T. durum, 325, 338, 344
tblastx, 539
Telomeric sequences, 327
Telosomic, 422
test, 538
tetraploid, 326
the Convention on Biological Diversity, 565
Threat information, 75
Threats, 75, 83, 84
- TILLING, 350, 588, 591–594
Tobacco, 512
Transcription, 595, 600–602, 604, 605
Transcriptome sequencing, 115–117, 123, 126
transcriptomics, 527
Transformation, 52, 588, 593, 594, 602, 604
transgenic, 51
transposable DNA element, 545
transposable elements (TE), 326, 329, 393, 406
tree genetic resources, 68, 72, 73
Tris-HCl, 331
type I phaseolin, 486
- U**
uniform, 296
Uniformity, and Stability, 34
unigene, 534, 538, 542, 548
uniparental inheritance, 495
upcoming, 534, 552
USDA- ARS, 501
Utility patents, 33
- V**
V. faba, 327
varieties, 684
variety identification, 33
vegetative propagation, 692
VIGS, 594
viticulture, 687
VV genome, 340
- W**
Wheat, 115, 117, 121–123, 325, 326, 406, 586–590, 594–600, 603, 605, 606
Wheat *Thinopyrum* (syn. *Agropyron*) Hybrids, 270
Wheat Rye Crossability, 263
wheat-*Ae. biuncialis* amphiploids, 269
wheat-*Th. ponticum* partial amphiploid, 272
Wheat-barley chromosome pairing, 259
wheat-rye addition lines, 264
Wheat/Barley Translocations, 261
Wheat/Rye Translocations, 264
Whole Chromosome Shotgun, 411
Wild *P. vulgaris*, 486
wild and the domesticated forms, 484
Wild *P. vulgaris*, 486
wild relatives, 483
- Y**
yield, 587, 588, 598, 601, 603, 605