Gaurav Sablok · Sunil Kumar
Saneyoshi Ueno · Jimmy Kuo
Claudio Varotto *Editors*

Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches

Springer

Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches

Gaurav Sablok • Sunil Kumar
Saneyoshi Ueno • Jimmy Kuo • Claudio Varotto
Editors

Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches

Springer

Editors
Gaurav Sablok
Biodiversity and Molecular Ecology
Research and Innovation Centre
Fondazione Edmund Mach
San Michele all'Adige, Italy

and

Plant Functional Biology and Climate Change Cluster, C3
University of Technology Sydney
Broadway, NSW, Australia

Saneyoshi Ueno
Department of Forest Genetics
Forestry and Forest Products Research Institute (FFPRI)
Tsukuba, Ibaraki, Japan

Claudio Varotto
Biodiversity and Molecular Ecology
Fondazione Edmund Mach
San Michele all'Adige, Italy

Sunil Kumar
Institute of Life Sciences
Nalco Square
Bhubaneswar, India

and

ICAR-NBAIM
Kushmaur, Mau (UP) India

Jimmy Kuo
Department of Planning and Research
National Museum of Marine Biology and Aquarium
Pingtung, Taiwan

ISBN 978-3-319-17156-2 ISBN 978-3-319-17157-9 (eBook)
DOI 10.1007/978-3-319-17157-9

Library of Congress Control Number: 2015942694

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Preface

Next Generation Sequencing has been leveraged primarily in the current era as a de facto for linking the biological hypothesis with the elucidation of the genes, biological pathways, and mechanistic evolution of certain traits and lineage specific evolutionary adaptations. Keeping in pace with the recent developments in the NGS technologies, several tools and techniques have been developed widely addressing questions of critical importance across the bacterial, fungal, and plant communities. Advances in the Understanding of the Biological Sciences using the Next Generation Sequencing is a compiled catalogue of such findings, where several NGS technologies ranging from the genomics, transcriptomics, metagenomics, single cell genomics, QTL, patho-genomics, and patho-transcriptomics have been applied to delineate the mystery of the associated mutations, biological pathway transitions, transcriptional fluxes, and patterns of host associated or adaptations to certain climatic conditions. The aims and scope of this book focus more on the biological underpinning to initiate the cross talks across the traits acquired or lost during the course of evolution. The structured framework of this volume provides the applicative point of view of the NGS technologies and demonstrates the conceptual way of linking the experimentation with the NGS technologies, to aid in researchers to place their biological hypothesis in a larger context.

This book would have not been accomplished without the support of my numerous colleagues, who have helped in editing the volume of the book, contributors, and motivational and organizational support of Prof. Peter Ralph, Executive Director, Plant Functional Biology and Climate Change Cluster, C3, University of Technology Sydney, Australia. This book would have not been possible without the encouragement and support of my wife Mrs. Namrata Sablok. Last but not least, thanks to Daniel, Jessica, and Ken (Springer Publishing) for making this volume finally published.

Broadway, NSW, Australia Gaurav Sablok, Ph.D.

Contents

Contributors

Akira Abe, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

James V. Anderson, Ph.D. Sunflower and Plant Biology Research Unit, USDA-ARS Red River Valley Agricultural Research Center, Bioscience Research Lab, Fargo, ND, USA

Liliana Cano, Ph.D. The Sainsbury Laboratory, Norwich Research Park, Norwich, Norfork, UK

Huansheng Cao, Ph.D. Department of Biological Sciences, Northern Illinois University, DeKalb, IL, USA

Wun S. Chao, Ph.D. Sunflower and Plant Biology Research Unit, USDA-ARS Red River Valley Agricultural Research Center, Bioscience Research Lab, Fargo, ND, USA

Bhupendra Chaudhary, Ph.D. School of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India

Münevver Doğramaci, Ph.D. Sunflower and Plant Biology Research Unit, USDA-ARS Red River Valley Agricultural Research Center, Bioscience Research Lab, Fargo, ND, USA

Vikash Kumar Dubey, Ph.D. Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India

Manu J. Dubin, Ph.D. Gergor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences, Vienna, Austria

Alex Ekstrom Department of Computer Science, Northern Illinois University, DeKalb, IL, USA

Mario A. Fares Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas, (CSIC-UPV), Ingeniero Fausto Elio, Valencia, Spain

Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland

Mobashar Hussain Urf Turabe Fazil, B.Sc., M.Sc., Ph.D. Dermatology and Skin Biology, Lee Kong Chian School of Medicine, Singapore

Rym Fekih, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Michael E. Foley, Ph.D. Sunflower and Plant Biology Research Unit, USDA-ARS Red River Valley Agricultural Research Center, Bioscience Research Lab, Fargo, ND, USA

Enrico Francia, Ph.D. Department of Life Sciences, University of Modena and Reggio Emilia, Reggio Emilia, Italy

CGR – Center for Genome Research, University of Modena and Reggio Emilia, Modena, Italy

Roger S. Greenwell Jr. , Ph.D. Biology Department, Worcester State University, Worcester, MA, USA

David Horvath, Ph.D. Sunflower and Plant Biology Research Unit, USDA-ARS Red River Valley Agricultural Research Center, Bioscience Research Lab, Fargo, ND, USA

Pär K. Ingvarsson, M.Sc. Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

Sophien Kamoun, Ph.D. The Sainsbury Laboratory, Norwich Research Park, Norwich, Norfork, UK

Hiroyuki Kanzaki, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Masaaki Kobayashi, Ph.D. School of Agriculture, Meiji University, Kawasaki, Kanagawa, Japan

CREST, JST, Kawaguchi, Saitama, Japan

Shunichi Kosugi, Ph.D. Kazusa DNA Research Institute, Kisarazu, Chiba, Japan

Ritesh Kumar, Ph.D. (pursuing) Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India

Sunil Kumar, M.Sc., I.P.R., M.C.A., Ph.D. Institute of Life Sciences, Nalco Square, Bhubaneswar, India

ICAR-NBAIM, Kushmaur, Mau (UP) India

Jimmy Kuo, Ph.D. Department of Planning and Research, National Museum of Marine Biology and Aquarium, Pingtung, Taiwan

Hideo Matsumura, Ph.D. Gene Research Center, Shinshu University, Ueda, Nagano, Japan

Satoshi Natsume Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Hajime Ohyanagi, Ph.D. School of Agriculture, Meiji University, Kawasaki, Kanagawa, Japan

CREST, JST, Kawaguchi, Saitama, Japan

National Institute of Genetics, Mishima, Shizuoka, Japan

Mitsubishi Space Software Co., Ltd., Tsukuba, Ibaraki, Japan

Dhananjay K. Pandey, M.Tech. School of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India

H.P. Pandey, Ph.D. Faculty of Science, Department of Biochemistry, Banaras Hindu University, Varanasi, Uttar Pradesh, India

Nicola Pecchioni, Ph.D. Department of Life Sciences, University of Modena and Reggio Emilia, Reggio Emilia, Italy

CGR – Center for Genome Research, University of Modena and Reggio Emilia, Modena, Italy

Alberto Policriti, Ph.D. Department of Mathematics and Computer Science, University of Udine, Udine, Italy

IGA - Institute of Applied Genomics, Parco Scientifico e Tecnologico "L. Danieli", Udine, Italy

Riccardo Rosselli, Ph.D. Department of Biology, University of Padova, Padova, Italy

Gaurav Sablok, Ph.D. Biodiversity and Molecular Ecology, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige, Italy

Plant Functional Biology and Climate Change Cluster, C3, University of Technology Sydney, Broadway, NSW, Australia

Hiromasa Saitoh, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Simone Scalabrin, Ph.D. Department of Mathematics and Computer Science, University of Udine, Udine, Italy

IGA Technology Services, Parco Scientifico e Tecnologico "L. Danieli", Udine, Italy

Douglas Scofield, Ph.D. Department of Genetics and Evolution, Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Shalini Singh, Ph.D. (pursuing) Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India

Andrea Squartini, Ph.D. Department of Agronomy Animals, Food, Natural Resources and Environment, DAFNAE, Legnaro, PD, Italy

Nathaniel R. Street, Ph.D. Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

Hiroki Takagi, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Muluneh Tamiru, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Ryohei Terauchi, Ph.D. Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Güzin Tombuloğlu, Ph.D. Vocational School of Medical Sciences, Fatih University, Istanbul, Turkey

Hüseyin Tombuloğlu, Ph.D. Department of Biology, Fatih University, Istanbul, Turkey

Saneyoshi Ueno, Ph.D. Department of Forest Genetics, Forestry and Forest Products, Research Institute (FFPRI), Tsukuba, Ibaraki, Japan

Claudio Varotto, M.Sc., Ph.D. Biodiversity and Molecular Ecology, Fondazione Edmund Mach, San Michele all'Adige, Italy

Jing Wang, M.Sc. Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

Hiroki Yaegashi Division of Genomics and Breeding, Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Kentaro Yano, Ph.D. School of Agriculture, Meiji University, Kawasaki, Kanagawa, Japan

CREST, JST, Kawaguchi, Saitama, Japan

Yanbin Yin, Ph.D. Department of Biological Sciences, Northern Illinois University, DeKalb, IL, USA

Kentaro Yoshida, Ph.D. The Sainsbury Laboratory, Norwich Research Park, Norwich, Norfork, UK

Chapter 1
Expression Analysis and Genome Annotations with RNA Sequencing

Masaaki Kobayashi, Hajime Ohyanagi, and Kentaro Yano

Abbreviations

CA	Correspondence analysis
HCL	Hierarchical clustering
NGS	Next-generation sequencing
SNP	Single nucleotide polymorphisms

Introduction

In order to detect genes on the basis of specific expression profiles, statistical approaches are frequently applied against large-scale expression data. Among the approaches, hierarchical clustering (HCL) (Eisen et al. 1998) has been widely used for expression data obtained from microarray and next-generation sequencing (NGS). The HCL method assists us in efficiently understanding the results via a graphical viewer consisting of dendrograms and heat maps. While this is a powerful

M. Kobayashi, Ph.D. • K. Yano, Ph.D. (✉)
School of Agriculture, Meiji University, Kawasaki, Kanagawa, Japan

CREST, JST, Kawaguchi, Saitama, Japan
e-mail: kyano@isc.meiji.ac.jp

H. Ohyanagi, Ph.D.
School of Agriculture, Meiji University, Kawasaki, Kanagawa, Japan

CREST, JST, Kawaguchi, Saitama, Japan

National Institute of Genetics, Mishima, Shizuoka, Japan

Mitsubishi Space Software Co., Ltd., Tsukuba, Ibaraki, Japan

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_1

tool, it requires substantial computer resources for analysis of large-scale datasets. To quickly detect specifically expressed genes from large-scale expression data, correspondence analysis (CA) (Greenacre 1993; Yano et al. 2006a), a multivariate analysis method, is useful. It does not require large computer resources, and genes can be projected into a two- or three-dimensional subspace, which helps us to understand expression profiles of each gene intuitively.

In this chapter, expression analysis and statistical methods with NGS technology are introduced. In recent years, the number of studies employing the NGS has increased markedly, as the sequencing costs for the use of NGS have decreased. Along with the reductions in cost, there have been considerable improvements in sequencing quality, read lengths, and data amounts from NGS. A large-scale mRNA-Seq analysis by NGS technology as well as microarray experiments provides genome-wide gene expression profiles. While microarray experiments require DNA probe design in advance, NGS technology requires neither complete genome sequences nor DNA probes. Therefore, genome-wide gene expression analysis with NGS technology has been widely applied for many organisms, for which microarray platforms have not been designed to date (Wang et al. 2009; Suzuki et al. 2013). Expression analysis with NGS is performed in multiple steps: (1) mRNA sequencing by NGS; (2) pre-processing of reads (obtained sequences); (3) mapping reads on a reference sequence(s); (4) assembling; and (5) expression profiling. Sequencing data from mRNA-Seq analysis are also available from public databases. The International Nucleotide Sequence Database Collaboration (INSDC), which consists of the NCBI Sequence Read Archive (Wheeler et al. 2008), the European Nucleotide Archive (ENA) of the EMBL-EBI (Leinonen et al. 2011), and the DDBJ Sequence Read Archive (DRA) (Kodama et al. 2012), store and distribute raw sequencing data produced by NGS. Genomic DNA sequence data used as reference genomes are available from web databases in each genome sequencing and annotation project, such as The *Arabidopsis* Information Resource (TAIR) for *Arabidopsis* (Lamesch et al. 2012), Rice Annotation Project Database (RAP-DB) for rice (Sakai et al. 2013), and the Sol Genomics Network (SGN) for tomato (The Tomato Genome Consortium 2012).

As well as differences in expression profiles among genes, the information on sequence polymorphisms including SNPs and SSRs among genes (including alleles, homologues) facilitates the understanding of biological functions of genes. Such sequence polymorphisms are identified by sequence analysis with NGS technology. The analysis methods for Genome-Seq are briefly described in this chapter.

Sequencing Strategy

In a fundamental manner, to detect genomic structural variations (mostly single nucleotide polymorphisms (SNPs) and small INDELs), the reference genome sequence and particular short-reads generated by Genome-Seq application are required, namely “genome resequencing.” When the goal is to profile the global gene expression levels, the reference genome sequence and the short-reads from mRNA-Seq will be needed (Wang et al. 2009).

In some way, the mRNA-Seq methods (not Genome-Seq methods) could also be appropriable to roughly predict structural variations in transcripts. The structural variations detected by using only mRNA-Seq data might be ambiguous, because these variations would be derived from both genome-level varieties and transcript-level varieties (e.g., difficulty in distinguishing the INDELs on genomes from the alternative splicing events on transcripts). Thus, the resultant variations might have mutual characteristics, and in this section, we do not discuss mRNA-Seq-based variation detection methods.

Genome Sequencing

Genome-Seq is the most basic application for genome resequencing in current NGS instruments. The most common platform, Illumina NGS, proposes three sequencing methods, namely single-end, paired-end, and mate-pair layouts, for their read libraries, depending on the purpose of DNA sequencing. Among these, the paired-end library is widely employed for genome resequencing. To comprehensively and accurately detect genome-wide structural variations and SNPs, a sufficient number of sequencing reads (sequencing depth resolution) is necessary. For most purposes, sequencing more than 20–30× short-reads against total genome size is preferable.

New experimental methods and biological methods are also proposed to efficiently predict genotypes with the Genome-Seq. Imputation method (Pei et al. 2008), allowing the effective deduction of missing genotypes. The RAD-Seq approach (Baird et al. 2008) facilitates detection of genome-wide SNPs among different genotypes (varieties, cultivars, and inbred lines) by confining the sequenced genome regions (see section "SNP Detecting from Genome Sequencing").

mRNA Sequencing

While the mRNA-Seq application requires particular lab techniques (poly-A selection for mRNA purification from total-RNA, reverse transcription into cDNAs), the instrumental rationale for mRNA-Seq is similar to that of Genome-Seq. As for reference-based mRNA-Seq application, illumina single-end or paired-end layouts are preferred. When the aim of sequencing is construction of unigenes, which are a non-redundant set of sequences (transcripts), and the strategy is reference-free de novo mRNA-Seq assembly, methods with relatively longer layout (GS FLX+, or illumina HiSeq2500 paired-end Rapid Mode) would be preferable for more efficient assembly outcomes.

In previous years, mRNA-Seq analysis remained an orientation-free methodology (e.g., knowledge of the strand origin of each mRNA molecule was unavailable). Now, a strand-specific mRNA-Seq procedure is available (Vivancos et al. 2010), confirming the biological significance of dubious anti-sense transcripts of particular genes that had been assumed to be experimental noise.

Read Mapping on Reference Sequences

For research based on high-throughput NGS analysis, short-read sequencing is generally performed because many reads are obtained by sequencing, as compared with the conventional long-read sequencing methods using Sanger technology. In this section, the pre-processing and mapping procedures in short-read sequencing are described.

Pre-processing

In order to correctly map reads on reference sequences, reads should be pre-processed. Pre-processing comprises multiple steps: (1) demultiplexing; (2) quality checking; (3) adapter trimming; and (4) quality controlling. These procedures are shown below.

(1) Demultiplexing

By taking advantage of the barcode (index) sequence system of Illumina instruments, multiple samples (e.g. genotypes, experimental lines) can be simultaneously sequenced on a single lane. A short series of nucleotides (tag), mostly 6-mer, are employed as a barcode sequence, which is inscribed within the PCR primer in advance. To discriminate reads according to the origins (samples) by referring the nucleotide pattern of the barcode sequence, a unique barcode sequence should be designed (determined) for each sample (DNA library) before sample preparation. Then the pooled sample from multiple DNA libraries is loaded on a single lane at once. Afterwards, the pooled sequencing data derived from multiple samples can be sorted into fractionated sequence data for each sample by scanning the barcode sequence pattern computationally. Generally, this procedure for categorization is known as "demultiplexing." Demultiplexing is simple to perform using a bioinformatics tool such as CASAVA (http://support.illumina.com/sequencing/sequencing_software/casava.ilmn).

(2) Quality checking

Quality checking of samples is recommended for proper read mapping. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) is a good tool for quality checking. It provides information on sequencing quality via statistical indices in HTML format. When sequencing qualities are considerably low, the sequencing method should be refined or improved.

(3) Adapter trimming

When inserted cDNA length is shorter than the length of the sequencing read, adapter sequences are consequently contained in read sequences. Artificial sequences disturb accurate mapping on reference sequences. Cutadapt is a suitable tool for trimming adapters, and in the case of mRNA-Seq, cutadapt is also to trim poly(A) sequences.

(4) Quality control

Low-quality bases in each read should be trimmed off, or the reads with low-quality bases should be discarded for further analysis. The FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) can be used for quality control. By trimming of low-quality regions in reads, high-quality reads can be obtained and correctly mapped on reference sequences.

Read Mapping

In the read mapping procedure, combinations of nucleotide types (genomic or mRNA sequences) of NGS reads and reference sequences must be taken into account. BWA (Burrows-Wheeler Aligner) (http://bio-bwa.sourceforge.net/) is a mapping tool that has been widely used in NGS data analysis, to align reads originating from genomic DNAs on genome sequences (Li and Durbin 2009). The mapping results of BWA are reported in a SAM (Sequence Alignment/Map) format. For mapping reads from mRNA-Seq on a reference genome sequence, large gaps (introns) in the sequence alignment should be taken into account. The combination of TopHat2 (http://tophat.cbcb.umd.edu/) and Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) allows the mapping of mRNA-Seq reads in consideration of splice junctions between exons. The results are also shown in the SAM format (Langmead and Salzberg 2012; Kim et al. 2013). For non-model plants, reference genome sequences are frequently unavailable. Unigene sequence data, which are predicted by assembling of ESTs and/or short-reads from mRNA-Seq (Yano et al. 2006b; Quackenbush et al. 2000; Habu et al. 2012), are often employed as reference sequences for mapping short-reads. For mRNA-Seq reads mapping onto mRNA sequences (e.g., unigenes, cDNA sequences), BWA and TopHat2/Bowtie2 are available.

SNP Detecting from Genome Sequencing

SNP information is fundamental data for comprehending functional differences between alleles, developing DNA markers for plant breeding and genome wide association studies, etc. (Yamamoto et al. 2010; Asamizu et al. 2012; Morris et al. 2013). NGS technology is also a powerful tool to mine genome-wide SNPs among genotypes (Austin et al. 2011; Arai-Kichise et al. 2011; Huang et al. 2012). For reads obtained from NGS, sequence mapping and alignment procedures onto genome DNA sequences are generally performed to detect SNPs. In this calculation, SNP data are summarized in SAM format files by using BWA or TopHat2/Bowtie2. Samtools and bcftools have been widely used to call SNP candidates and estimate allele frequencies from SAM format files. The results obtained with samtools and bcftools are described in VCF (Variant Call Format) files (Li et al. 2009).

In recent years, a new effective and efficient method, RAD-Seq (restriction-site associated DNA sequencing), has been proposed to comprehensively mine SNPs among multiple genotypes (Baird et al. 2008). The RAD-Seq approach has been applied in comprehensive studies, for example, evolutionary analysis and QTL mapping (Etter et al. 2011; Houston et al. 2012).

Stacks is a useful tool for calling SNPs by using RAD-Seq data (Catchen et al. 2011). Stacks directly aligns and compares short-read sequences, which are predicted as DNA fragments from the same genomic region in the chromosome, among different genotypes. Therefore, with or without the reference genome sequence data, Stacks can mine SNPs with RAD-Seq data. When there are no reference genome sequence data, researchers can import fasta or fastaq format files into Stacks to align reads. Otherwise, researchers can import SAM format files obtained by mapping when reference genome sequence data are available. This tool provides a list of SNP candidates in TSV or Excel format.

Digital Gene Expression Profiling by mRNA-Seq

Microarray analysis was the primary method of obtaining global gene expression profiles in the early 2000s. While microarray techniques have been predominantly employed for gene expression analyses, particularly for well-annotated model species, the current explosive growth of NGS technology has made it obsolete. Digital gene expression profiling by mRNA-Seq has been gaining popularity and is recognized to have advantageous points over microarray technology (Wang et al. 2009). The principle of mRNA-Seq is simple; it estimates absolute gene expression levels by counting the number of mRNA molecules on each gene model (alleles). In this section, the procedures for mRNA-Seq, particularly with Illumina instruments, will be reviewed together with a discussion of gene expression level correction.

Mapping

Once the pre-processing steps (see section “Pre-processing”) have been completed, the rest of the mRNA-Seq reads are to be mapped on the reference genome sequences (see section “Read Mapping”). In contrast to the simpler whole-genome sequence samples, the mapping procedure of mRNA-Seq reads on reference genome sequences should consider exon–intron structures. This mapping process can be optimally performed by the combination of TopHat2 (Kim et al. 2013) and Bowtie2 (Langmead and Salzberg 2012) (see section “Read Mapping”). Reads that have located repeatedly on the reference would also be handled properly with this software (see section “Expression Level Estimation and Correction”).

Assembly

After estimating the original location of each read, gene models should be reconstructed from a body of mapped reads for the following gene expression profiling step. Cufflinks combined with TopHat2 (Trapnell et al. 2012) can perform this assembly step. The reference gene annotations can also be transmitted to Cufflinks together with the reference genome DNA sequences, while novel gene models can also be predicted by referring to the assembly outcomes of short-reads. This capability in novel gene detection is one of the advantages of mRNA-Seq technology over microarray analyses, which require probes to be designed in advance. Cufflinks also calculates the FPKM (*f*ragments *p*er *k*ilobase of exon per *m*illion mapped reads) of each gene model, which is a widely used normalized intensity for gene expression levels (see section "Expression Level Estimation and Correction").

Expression Level Estimation and Correction

Assembling results with total reads from different samples (e.g., genotypes, developmental stages, organs, and treatments) give us clues to understand the differences in terms of biological aspects or behaviors (such as response to stress) among samples. Merging information on gene models and FPKM and detecting differentially expressed gene models can be conducted by cuffcompare and cuffdiff (subprograms implemented in Cufflinks) (Trapnell et al. 2012).

In the profiling process, the tag count of each gene should be normalized against both exon length of each gene and total number of mapped reads, as exon lengths should differ among genes, and the total number of tags should also vary across multiple samples. Here, RPKM stands for *r*eads *p*er *k*ilobase of exon model per *m*illion mapped reads (Mortazavi et al. 2008), and reflects the molar concentration of a transcript in each sample by normalizing for RNA length and for total read count in the measurement. This facilitates transparent comparisons of transcript levels both within and between samples. FPKM (Trapnell et al. 2012), which stands for *f*ragments *p*er *k*ilobase of exon per *m*illion mapped reads, is a similar value. While RPKM simply sums the applicable reads as the numerator and denominator of the formula, FPKM takes account of each of the "paired reads" in order to avoid overestimation of gene expression, as both members of each pair must be derived from an identical mRNA molecule (Trapnell et al. 2012).

Some of the reads could be mapped on the reference genome multiple times due to sequence repeats and homology within the genome DNA sequences. In the case of Cufflinks, it evenly divides each multi-mapped read to all of the positions it maps to. For instance, a read mapped to 4 positions will count as 25 % of a read at each position by default. Cufflinks also offers more sophisticated multi-map correction by the "rescue method" (Mortazavi et al. 2008) as an optional extra.

Statistical Methods for Gene Expression Data

To date, numerous statistical methods have been proposed and used in gene expression analysis. Among these, several statistical methods are described here. These include HCL, self-organizing maps (SOMs), *k*-means, and CA.

Hierarchical Clustering

HCL is one of the methods that has been widely used in gene expression analysis. Eisen et al. (1998) provided freely available GUI software "CLUSTER" for execution of HCL and "TreeView" for graphically browsing the results (http://rana.lbl.gov/EisenSoftware.htm). Using the software "CLUSTER," users can perform not only HCL but also SOMs, *k*-means clustering, and principal component analysis (PCA). As the original software is executed only on Windows computers, a new modified version (CLUSTER 3.0) which is available for Windows, Mac OS X, and Linux/Unix platforms has been provided by the University of Tokyo (http://bonsai.hgc.jp/~mdehoon/software/cluster/) (de Hoon et al. 2004). CLUSTER 3.0 can be used both as a GUI program and as a command line program.

In general, genes with similar expression profiles are clustered by HCL analysis. The similarities in expression profiles among genes can be browsed using a graphical viewer showing a dendrogram. As genes located in the same subtree in the dendrogram should have similar expression profiles, genes with similar or specific expression profiles can be quickly detected using the graphical viewer and dendrogram.

For large-scale expression data, HCL requires large-scale computational resources (Yano et al. 2006a). When calculations for large-scale expression data cannot be performed by HCL with a practical turnaround time, eliminating genes with small fold changes in expression levels for further analysis allows us to perform HCL within a short amount of time. Although this approach makes the calculation executable, the eliminated genes may contain some related to traits of interest. For calculations against large-scale expression data by HCL, elimination must be carefully considered. Or, instead of HCL, multivariate analysis methods including SOMs and CA are effective to quickly execute large-scale expression data analysis.

Multivariate Analysis Methods; Principal Component Analysis (PCA) and Correspondence Analysis (CA)

Multivariate analysis methods can handle large-scale data, even with an entry-level workstation or a personal computer used in an ordinary laboratory. Therefore, PCA and CA have often been applied to large-scale omics data analysis (Yano et al. 2006a; Pomeroy et al. 2002; Hirai et al. 2004). Both methods allow us to summarize a data matrix that is originally high-dimensional (row [gene] and column [sample])

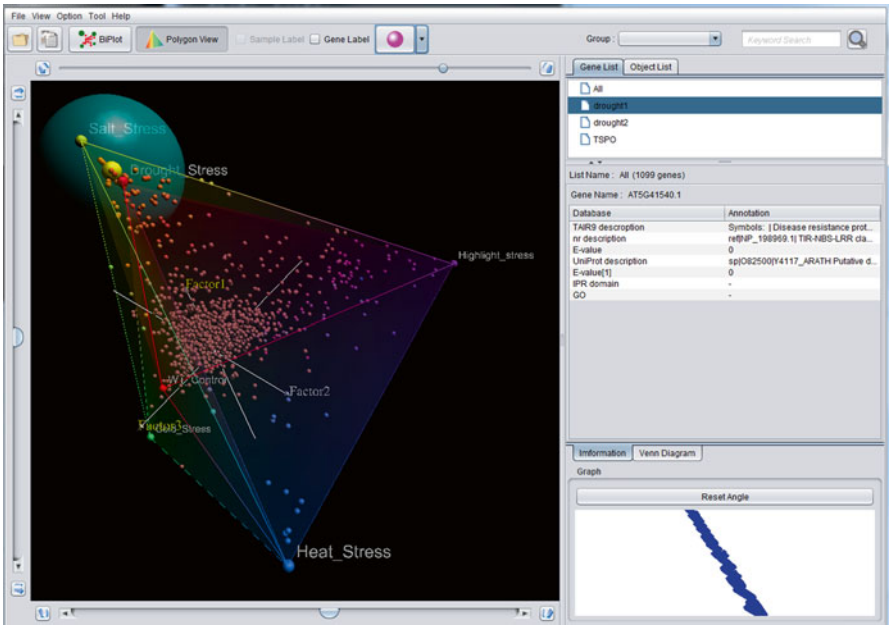

Fig. 1.1 An interface of GUI software "CA_Plot_Viewer." By using "CA_Plot_Viewer," CA can be easily and quickly executed in Windows, Macintosh, and Linux PCs. Users can efficiently search genes according to expression profiles and functional annotations in the graphical viewer. Various annotations can be also imported into the software

with a low-dimensional projection. Each gene and sample is given coordinates in the low-dimensional space. With the coordinates, genes and/or samples can be projected into a 2- or 3-dimensional subspace in order to graphically show and understand the similarities and/or specificities of expression profiles among genes. When the aim of the gene expression analysis is to categorize genes according to the similarities in expression profiles, CA is more effective than PCA (Greenacre 1993; Yano et al. 2006a). When CA is applied to omics data, all values in the matrix must be zero or positive. Therefore, much attention should be paid in the normalization methods before CA. For example, a log-transformed method may contain negative values, which cannot be handled in CA.

PCA and CA can be performed with some pieces of computer software. For example, R packages "mass" and "ca" help us to execute CA. To globally interpret the results obtained from CA, a graphical and intuitive viewer is also required, as large-scale expression analysis frequently contains more than tens of thousands of genes. This viewing software should allow graphical adjustments such as rotation, zooming in and out, and panning of the image and searches for genes with functional annotations (descriptions) (Hamada et al. 2011; Manickavelu et al. 2012; Nishida et al. 2012). For this purpose, a GUI tool and viewer for CA calculation "CA Plot Viewer" has been developed and distributed (http://bioinf.mind.meiji.ac.jp/lab/) (Fig. 1.1).

Self-Organizing Maps and k*-Means Clustering*

SOMs represent a useful clustering method. They can be quickly performed for large-scale expression data. Genes are classified according to expression patterns, and given coordinates. Genes in the same cluster show similar expression profiles. An R package "som" allows for two-dimensional SOM analysis and graphically shows the results (http://cran.r-project.org/web/packages/som/index.html). In SOM analysis, the user must decide the number of clusters before analysis.

The clustering method *k*-means has been also widely used to analyze gene expression data. Similarly to SOM, *k*-means clustering can be performed quickly for large-scale expression data.

References

Arai-Kichise Y, Shiwa Y, Nagasaki H, Ebana K, Yoshikawa H, Yano M et al (2011) Discovery of genome-wide DNA polymorphisms in a landrace cultivar of *Japonica* rice by whole-genome sequencing. Plant Cell Physiol 52(2):274–282. PubMed PMID: 21258067; PubMed Central PMCID: PMC3037082

Asamizu E, Shirasawa K, Hirakawa H, Sato S, Tabata S, Yano K et al (2012) Mapping of micro-tom BAC-End sequences to the reference tomato genome reveals possible genome rearrangements and polymorphisms. Int J Plant Genomics 2012:437026. PubMed PMID: 23227037; PubMed Central PMCID: PMC3514829

Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, Bonetta D et al (2011) Next-generation mapping of *Arabidopsis* genes. Plant J 67(4):7157–7125

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10):e3376. PubMed PMID: 18852878; PubMed Central PMCID: PMC2557064

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping Loci *de novo* from short-read sequences. G3 1(3):171–182. PubMed PMID: 22384329; PubMed Central PMCID: PMC3276136

de Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20(9):1453–1454

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95(25):14863–14868. PubMed PMID: 9843981; PubMed Central PMCID: PMC24541

Etter PD, Bassham S, Hohenlohe PA, Johnson EA, Cresko WA (2011) SNP discovery and genotyping for evolutionary genetics using RAD sequencing. Methods Mol Biol 772:157–178. PubMed PMID: 22065437; PubMed Central PMCID: PMC3658458

Greenacre MJ (1993) Correspondence analysis in practice. Academic, London

Habu T, Yamane H, Igarashi K, Hamada K, Yano K, Tao R (2012) 454-Pyrosequencing of the Transcriptome in leaf and flower buds of Japanese apricot (*Prunus mume* Sieb. et Zucc.) at different dormant stages. J Jpn Soc Hortic Sci 81(3):239–250. WOS: 000306721500003

Hamada K, Hongo K, Suwabe K, Shimizu A, Nagayama T, Abe R et al (2011) Oryza Express: an integrated database of gene expression networks and omics annotations in rice. Plant Cell Physiol 52(2):220–229. PubMed PMID: 21186175; PubMed Central PMCID: PMC3037078

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M et al (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 101(27):10205–10210. PubMed PMID: 15199185; PubMed Central PMCID: PMC454188

Houston RD, Davey JW, Bishop SC, Lowe NR, Mota-Velasco JC, Hamilton A et al (2012) Characterisation of QTL-linked and genome-wide restriction site-associated DNA (RAD) markers in farmed Atlantic salmon. BMC Genomics 13:244. PubMed PMID: 22702806; PubMed Central PMCID: PMC3520118

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q et al (2012) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44(1):32–39

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14(4):R36

Kodama Y, Shumway M, Leinonen R (2012) International nucleotide sequence database collaboration. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 40(Database issue):D54–D56. PubMed PMID: 22009675; PubMed Central PMCID: PMC3245110

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R et al (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40(Database issue):D1202–D1210. PubMed PMID: 22140109; PubMed Central PMCID: PMC3245047

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359. PubMed PMID: 22388286; PubMed Central PMCID: PMC3322381

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y et al (2011) The European nucleotide archive. Nucleic Acids Res 39(Database issue):D28–D31. PubMed PMID: 20972220; PubMed Central PMCID: PMC3013801

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. PubMed PMID: 19451168; PubMed Central PMCID: PMC2705234

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/map format and SAM tools. Bioinformatics 25(16):2078–2079. PubMed PMID: 19505943; PubMed Central PMCID: PMC2723002

Manickavelu A, Kawaura K, Oishi K, Shin IT, Kohara Y, Yahiaoui N et al (2012) Comprehensive functional analyses of expressed sequence tags in common wheat (*Triticum aestivum*). DNA Res 19(2):165–177. PubMed PMID: 22334568; PubMed Central PMCID: PMC3325080

Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD et al (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci USA 110(2):453–458. PubMed PMID: 23267105; PubMed Central PMCID: PMC3545811

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5(7):621–628. PubMed

Nishida H, Abe R, Nagayama T, Yano K (2012) Genome signature difference between *Deinococcus radiodurans* and *Thermus thermophilus*. Int J Evol Biol 2012:205274. PubMed PMID: 22500246; PubMed Central PMCID: PMC3303625

Pei YF, Li J, Zhang L, Papasian CJ, Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS One 3(10):e3551. PubMed PMID: 18958166; PubMed Central PMCID: PMC2569208

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME et al (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415(6870):436–442

Quackenbush J, Liang F, Holt I, Pertea G, Upton J (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res 28(1):141–145. PubMed PMID: 10592205; PubMed Central PMCID: PMC102391

Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y et al (2013) Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 54(2):e6. PubMed PMID: 23299411; PubMed Central PMCID: PMC3583025

Suzuki T, Igarashi K, Dohra H, Someya T, Takano T, Harada K et al (2013) A new omics data resource of pleurocybellaporrigens for gene discovery. PLoS One 8(7):e69681. PubMed PMID: 23936076; PubMed Central PMCID: PMC3720577

The Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485(7400):635–641. PubMed PMID: 22660326; PubMed Central PMCID: PMC3378239

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578. PubMed PMID: 22383036; PubMed Central PMCID: PMC3334321

Vivancos AP, Guell M, Dohm JC, Serrano L, Himmelbauer H (2010) Strand-specific deep sequencing of the transcriptome. Genome Res 20(7):989–999. PubMed PMID: 20519413; PubMed Central PMCID: PMC2892100

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. PubMed PMID: 19015660; PubMed Central PMCID: PMC2949280

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V et al (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 36(Database issue):D13–D21. PubMed PMID: 18045790

Yamamoto T, Nagasaki H, Yonemaru J, Ebana K, Nakajima M, Shibaya T et al (2010) Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11:267. PubMed PMID: 20423466; PubMed Central PMCID: PMC2874813

Yano K, Imai K, Shimizu A, Hanashita T (2006a) A new method for gene discovery in large-scale microarray data. Nucleic Acids Res. 34(5):1532-9. PubMed PMID: 16537840; PubMed Central PMCID: PMC1401514

Yano K, Watanabe M, Yamamoto N, Tsugane T, Aoki K, Sakurai N, Shibata D (2006b) MiBASE: a database of a miniature tomato cultivar Micro-Tom. Plant Biotechnol 23(2):195–198

Chapter 2
The Application of Next Generation Sequencing Techniques to Plant Epigenomics

Manu J. Dubin

Introduction

It has long been known that an organism's appearance and phenotype are primarily determined by the genetic information encoded at the DNA sequence level within its genome. However over the last decades it has become increasingly apparent that an additional non-mendelian or "epigenetic" layer of information exists that influences gene expression, cell-type specification, phenotype and environmental responses (Bird 1984; Jenuwein and Allis 2001; Madlung and Comai 2004; Bossdorf et al. 2008). To a large extent this epigenetic information controls the behaviour of genes other genetic elements by altering the chromatin environment around them. This occurs by a variety of processes including DNA methylation, histone modifications such as acetylation, methylation, ubiquitination phosphorylation, and chromatin binding proteins (Taverna et al. 2007; Bannister and Kouzarides 2011).

Plants are sessile organisms that must adapt their development and metabolism to the prevailing environmental conditions (Kalisz and Kramer 2008), and epigenetic mechanisms appear to play prominent role in these processes (Matzke et al. 2009; Zhang et al. 2013). These include cell-type specification and regulation of gene expression at different developmental stages and during developmental transitions such as flowering (Sung and Amasino 2004; Costa and Shaw 2007) and maintaining gene expression through cell division (Jenuwein and Allis 2001), environmental responses, and environmental adaptation and stress responses (Bossdorf et al. 2008; Sani et al. 2013). Epigenetic processes also play a role in the imprinting of gene expression (Autran et al. 2011). The maintenance of genome integrity including DNA damage repair, silencing of transposons, and defense

M.J. Dubin, Ph.D. (✉)
Gregor Mendel Institute of Molecular Plant Biology, Austrian Academy of Sciences,
Dr. Bohr-Gasse 3, Vienna 1030, Austria
e-mail: manu.dubin@gmi.oeaw.ac.at

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_2

responses are also regulated by epigenetic mechanisms (Martienssen and Colot 2001; Tompa et al. 2002; Takeda et al. 2004; Dowen et al. 2012). Although somewhat controversial, there is evidence that epigenetic processes are involved in the transgenerational maintenance of phenotypes in the apparent absence of DNA sequence variation (Jacobsen and Meyerowitz 1997; Cubas et al. 1999; Paszkowski and Grossniklaus 2011).

Major Findings in Plants

The first demonstration of epigenetic phenomena was the Nobel prize-winning work by Barbara McClintock in the 1940s and 1950s showing that transposons could move (paramutate) and affect the expression of genes (McClintock 1950, 1956). Since then many other breakthroughs in the epigenetics field have come from studies in plants including the role of RNAs in disease resistance and transgene silencing (English et al. 1996). The occurrence of spontaneous epimutants that stably maintain differences in gene expression over multiple generation, even in the absence of DNA sequence variation (Jacobsen and Meyerowitz 1997; Cubas et al. 1999) and the role of small RNA's in directing de novo DNA methylation and silencing (Kanno et al. 2005).

Development of Genome-Wide Approaches

The field of epigenomics research was initiated by the development of microarray-based methods that allowed DNA methylation and chromatin modifications to be assayed on a genome-wide scale. These methods included ChIP-on-CHIP where chromatin is immunoprecipitated, the DNA is purified, amplified, and labeled and then hybridized to a microarray chip (Bernstein et al. 2002; Weinmann et al. 2002). An analogous method for looking at DNA methylation was MeDIP-ChIP (Methylated DNA immunoprecipitation followed by hybridization to microarray chip). Here DNA is purified and sheared, then methylated fraction of the genome is immunoprecipitated with an antibody against methylated DNA. The precipitated DNA is then amplified, labeled, and hybridized to a microarray (Weber et al. 2005). Variations of this protocol exist, including the use of the immobilized recombinant methylated DNA binding domain (MBD) to precipitate the DNA instead of an antibody (Ballestar et al. 2003). Other approaches include comparing the hybridization efficiency to a microarray of DNA digested with methylation sensitive restriction enzymes relative to DNA digested with a methylation insensitive enzyme (Schumacher et al. 2006). A related approach is to bisulfite treat the DNA (so all unmethylated cytosines are deanimated to uracil) and compare its hybridization efficiency to non-treated DNA (Gitan et al. 2002).

The use of these microarray-based approaches allowed the characterization of the enzymes involved in heterochromatin formation and transposon silencing (Tompa et al. 2002) and further clarified the role of small RNAs in targeting DNA methylation to transposons and other repetitive elements (Lippman et al. 2004; Teixeira et al. 2009). Using ChIP-on-ChIP the genome-wide distribution of different histone modifications was also investigated (Martienssen et al. 2005). Microarrays have also been used to determine the abundance of small RNAs in plants (Boccara et al. 2007). Further studies have been used to define the major chromatin states in *Arabidopsis* (Roudier et al. 2011) and to identify the genome-wide binding sites of chromatin binding proteins (Turck et al. 2007; Benhamed et al. 2008) and transcription factors (Oh et al. 2009; Zheng et al. 2009).

These methods generally require the use of tiling arrays (which short overlapping probes covering the entire genome), which are only available for a few species. For this reason in plants these methods have been largely limited to *Arabidopsis*, with a few studies in rice (Li et al. 2008c) and maize (Eichten et al. 2011). Even within a species microarray-based approaches are sensitive to the genetic background, and although they have been used to look at variation between different natural accessions, this requires careful experimental design to correct for copy number variants (Vaughn et al. 2007; Moghaddam et al. 2011).

More recently the development of next generation sequencing based approaches (Bentley 2006) has overcome some of the limitations of earlier microarray-based methods. In the case of ChIP studies the precipitated DNA is sequenced using next generation sequencing to determine the enriched regions, thus eliminating the need for a microarray (Barski et al. 2007). Advantages of this approach include improvements in quantitation, sensitivity, and the ability to discriminate between similar DNA sequences (i.e., potentially allowing allele specific investigations (Ho et al. 2011). Significantly the advent of these methods allows for experiments to be preformed in a wider range of species or cultivars for which a suitable tiling microarray is not available.

Bisulfite-Seq Experiments

Sanger sequencing of bisulfite-converted DNA allows the determination of methylation status of DNA at loci of interest at the single nucleotide level (Clark et al. 1994). This is particularly useful in plants where DNA can be methylated in the CG, CHG, and CHH contexts (where H is any nucleotide except guanine) as it allows the pathways to be dissected individually (Wassenegger et al. 1994; Ronemus et al. 1996; Lindroth et al. 2001). While Sanger bisulfite sequencing is limited to a handful of loci, the advent of next generation sequencing allowed the determination of the methylation status of the entire genome at the single nucleotide level (Cokus et al. 2008; Lister et al. 2008). These methods together with small-RNA sequencing have been fundamental for driving forwards the field of plant epigenomics research. Analysis of methylation patterns over many generations has shown that, although

largely stable, they can evolve in the absence of obvious DNA sequence variation and cause stably inherited changes in gene expression (Becker et al. 2011; Schmitz et al. 2011). In a series of papers it was shown that active demethylation of both the male gametophyte and maternal companion cell activates the expression of transposons in these cells, which produce small 24 nucleotide RNAs that appear to move to the germline where they reinforce silencing of transposons, thus ensuring genomic integrity during reproduction (Hsieh et al. 2009; Calarco et al. 2012; Ibarra et al. 2012). Small RNAs and DNA methylation have also appeared to have roles in parental imprinting of gene expression in early plant development (Autran et al. 2011; Vu et al. 2013). Other studies have looked at the effect of natural genetic variation on DNA methylation patterns (Schmitz et al. 2013b). DNA methylation patterns have also been described in a variety of plant species including *Brachypodium*, rice, soybean, poplar, and green algae (Feng et al. 2010; Rodrigues et al. 2013; Schmitz et al. 2013a; Takuno and Gaut 2013). A good source of information and protocols for plant epigenomics is the Epigenomics of Plants International Consortium (EPIC, www.plant-epigenome.org).

Methods

General Considerations

Virtually all NGS based epigenomics methods require a reference genome sequence against which the short reads can be aligned. In the absence of a reference it may be feasible to use that from a closely related species or sub-species (He et al. 2010).

At the time of writing the Illumina sequencing-by-synthesis (www.illumina.com) and to a lesser extent the SOLiD platform from Applied Biosystems (www.appliedbiosystems.com) have been used in the vast majority of published epigenomics studies. However NGS technologies are evolving at a rapid pace and other platforms may also be worth considering. Alternatives include the Ion torrent platform (www.iontorrent.com), which may be suitable for small genomes or where low coverage is acceptable. Likewise the 454 (www.454.com) and PacBio (www.pacificbiosystems.com) systems offer lower coverage but give longer read lengths, which may be advantageous in some situations. The PacBio platform is capable of directly detecting DNA methylation (Flusberg et al. 2010), although at the time of writing this technique appears not to be widely utilized.

It is highly advisable to discuss the experimental design with a Statistician and Bioinformation prior to initiating an epigenomics study in order to ensure that a data set that will be able to answer the question asked may be obtained. Points to consider include are inclusion of biological replicates, the depth (number of reads) and type of sequencing (Single or paired end, length) required. Thought should also be given to the availability of expertise and computational resources for downstream data processing.

Techniques for Assaying DNA Methylation in Plants Using NGS

DNA Methylation

The most widely used methods for determining DNA methylation using NGS approaches are methylated DNA immunoprecipitation and sequencing (MeDIP-Seq; Weber et al. 2005) and whole-genome bisulfite sequencing (Cokus et al. 2008; Lister et al. 2008).

In MeDIP-Seq DNA is purified and sheared by sonication or a similar method and then the methylated fraction is affinity purified using an antibody against methylated cytosine (Cortijo et al. 2014). The precipitated DNA is purified and used to prepare a library for illumina sequencing. When the reads from this precipitated library are mapped to the genome, they should be methylated regions of the genome.

Advantages of MeDIP-Seq include the relative simplicity of the technique, the modest amounts of starting material required, lower sequencing costs (50 bp single end vs 100 bp paired end typically used for BS-Seq), and easier alignment to the genome. The primary downsides are the inability to discriminate between different methylation context, difficulties in quantifying absolute levels of methylation and limited resolution (e.g., down to ~100 bp).

BS-Seq also starts with purified and sheared DNA. Standard NGS library preparation is carried out up until the adapter ligation step (typically using a commercial library preparation kit). Adapters where all the cytosines are methylated are then ligated and the samples are then denatured with heat or basic conditions and incubated with sodium bisulfite. This has the effect of converting all the non-methylated cytosines in the DNA sample to uracil (Li et al. 2013a). The bisulfite-converted library is then amplified using a polymerase that can convert uracil to thymine such as *PfuTurbo Cx* (Agilent Technologies, USA) or HiFi Uracil+ (KAPA, USA). The library is then sequenced in normal manner.

The use of BS-Seq is particularly advantageous in plants as its single nucleotide resolution allows the discrimination of CG, CHG, and CHH methylation. Limitations include the relatively large amounts of starting material required (as significant amounts of DNA are degraded during the bisulfite conversion process) and difficulties in mapping bisulfite-converted reads due to the decreased sequence complexity (Krueger et al. 2012; Hardcastle 2013). The use of long paired end reads can be used to increase both the mapping accuracy and sequence coverage. Detailed protocols can be found in (Li et al. 2013a) or on the website of the Epigenomics of Plants International Consortium (www.plant-epigenome.org).

ChIP-Seq

Chromatin immunoprecipitation can be used to determine the genomic localization of a wide range of chromatin associated proteins including transcription factors, chromatin binding proteins and modifying proteins as well as different histone

modifications and different histone variants (Kaufmann et al. 2010; Luo et al. 2012; Wollmann et al. 2012; Zemach et al. 2013). In particular the epigenomics field has benefited immensely from the commercial availability of high affinity antibodies against a wide array of histone modifications, which has allowed their genomic distribution to be determined (Luo et al. 2012).

While there are a multitude of variations of chromatin immunoprecipitation protocols, in plants most assays begin with formaldehyde crosslinking of the chromatin using vacuum infiltration (Kaufmann et al. 2010). Following this, nuclei are purified and then sonicated to shear the chromatin to approximately 250–500 bp. The chromatin is then incubated with an antibody against the antigen of interest (e.g., a modified histone) and the chromatin-antibody complex captured with a suitable affinity matrix (e.g., protein-A sepharose). After several wash steps, the purified chromatin is eluted from the affinity matrix under denaturing conditions, the formaldehyde cross-links reversed with heat and the DNA component of the chromatin purified.

ChIP experiments typically also include controls for normalization purposes. No-Antibody or pre-immune serum controls are common in conventional ChIP experiments but are of limited use in ChIP-Seq as they are unlikely to contain enough DNA for successful library construction. Instead a library containing a fraction of input material used for the immunoprecipitation is typically used. In some cases, e.g. when performing ChIP on modified histone, a control ChIP on the unmodified form of the histone can be performed in parallel and used for normalization purposes.

Many variations of ChIP-Seq are possible, for example if targeting a transcription factor or chromatin binding protein a longer range cross linker such as EGS (ethylene glycol bis-succinimidylsucciniate) can be used in addition to formaldehyde (Zeng et al. 2006). In other cases, such as when looking at histone modifications native ChIP may instead be employed. In this case no crosslinker is used, instead nuclei are directly purified from the tissue and then the chromatin fragmented using micrococcal nuclease instead of sonication (O'Neill and Turner 2003; Gilfillan et al. 2012). Important considerations in ChIP-Seq experiments include the specificity of the antibody and the stringency of the washing steps. Unlike other ChIP experiments, in ChIP-Seq it is important that the affinity matrix is not blocked with any competitor DNA (such as salmon-sperm DNA) as this may contaminate the ChIP-Seq library.

Prior to library construction, the purified DNA can be tested with endpoint or quantitative PCR to confirm that genomic regions known to associate with the antigen used are enriched. If this is successful and enough DNA is present (generally a minimum of 5 ng is required), a library can be prepared from the remaining material, preferably using methods that minimize sample loss. Often this is done by using ChIP-Seq library preparation kit that is optimized to work with low amounts of starting material. Detailed ChIP-seq protocols can be found in (Kaufmann et al. 2010; Zhu et al. 2012) while detailed protocols for troubleshooting and optimizing different steps of the ChIP protocol have been described in (Haring et al. 2007). For ChIP-seq the reads only need to be long enough to map accurately to the genome, so generally 36 or 50 bp single end sequencing is used

Nucleosome Positioning

The positioning of nucleosomes has a role in the specification of promoter regions, transcription start sites, and enhancers (Jiang and Pugh 2009; Chodavarapu et al. 2010; Rada-Iglesias et al. 2011). The expression levels of genes are also typically reflected in the pattern of the nucleosomes with higher nucleosome density and well-positioned nucleosomes on transcribed exons (Chodavarapu et al. 2010).

Although nucleosome positioning can be inferred from ChIP-Seq data, higher resolution can be obtained using micrococcal nuclease digestion followed by NGS (MNase-Seq). In this approach nuclei are first purified from tissue and then nucleosomes are solubilized by digestion with micrococcal nuclease which digests the linker DNA between adjacent nucleosomes, but not the DNA wrapped around the nucleosomes themselves (Wei et al. 2012). The nucleosome bound DNA is then purified and used for NGS library preparation. Normally fairly short reads (e.g., 36–50 bp) are sufficient to accurately map the reads to the genome. Although not strictly required, paired end reads will allow the determination of nucleosome position with higher accuracy.

Small RNA-Seq

A wide variety of small RNA species are produced in plants that are involved in the modulation of gene expression and the silencing of transposons and other repetitive sequences. These include micro-RNAs, trans-acting small RNAs, and 21 and 24 nucleotide small RNAs (Parent et al. 2012). 24 nucleotide small RNAs are produced by the RNA dependent DNA methylation (RdDM) pathway where transposons and other repetitive sequences have been transcribed by DNA pol IV. These transcripts are converted to double stranded RNA by an RNA dependent RNA polymerase RDR1 and then diced into 24 nucleotide small RNAs by the Dicer-like 3 enzyme. These small RNAs then guide the domains rearranged methyltransferases (DRM1/2) to sites with homology in a process that involves DNA pol V (Matzke et al. 2009).

Although 24 nucleotide small RNAs are primarily of interest from the epigenomics perspective, small RNA-sequencing experiments will also detect 21 nucleotide small RNAs, microRNAs (miRNAs), and trans-acting small RNAs (ta-siRNAs). These are involved in gene post-transcriptional gene silencing but may also be of interest to the researcher.

Small RNA-Seq libraries are prepared from total RNA extracted from fresh or frozen tissue using trizol or a commercial RNA purification kit. RNAs in the 18–30 nucleotide size range are then purified by polyacrylamide gel electrophoresis. Adenylated 3′ and 5′ adapters are sequentially ligated to the RNA and reverse transcription is performed. The library is then amplified by PCR and purified (Lu et al. 2007; Hafner et al. 2008; Chen et al. 2012; Linsen and Cuppen 2012). Due to their small size, short single end reads of 36 or 50 bp are sufficient for small RNA-Seq.

PART3: Analysis of Next Generation Sequencing Data for Epigenomics

Computing Requirements for Data Analysis

Due to the large amounts of data produced by NGS experiments a fairly powerful computer is generally required. Many programs used in NGS analysis are designed to run in a UNIX command-line environment (e.g., Linux or OSX). They can also be run on other operating systems using a UNIX emulator such as Cygwin (www.cygwin.com). For downstream processing and statistical analysis many specialized packages are available for the R statistical environment (www.r-project.org/) via the Bioconductor project (www.bioconductor.org/).

An alternative for those with limited computing resources and/or limited command line expertise is the process the data on a local or remote server via a GUI (graphical user interphase) environment such as Galaxy (www.galaxyproject.org; Blankenberg et al. 2010) which integrates many of the command line tools described here.

Quality Control

On receiving raw NGS data the first step is to check the raw data for obvious problems, This can be done by examining the quality control report if one is provided by the sequencing facility. If none is provided, it can be generated using (for example) the FASTQC pipeline software (www.bioinformatics.babraham.ac.uk/projects/fastqc/) which is also available in the Galaxy environment (www.galaxyproject.org). Points to consider include fraction of reads that can be mapped to genome, fraction of contaminating sequences and biases in nucleotide frequency, duplication levels and base-call qualities. It is worthwhile to keep in mind that some epigenomics techniques by their nature generate biases that are incorrectly flagged by the FastQC pipeline. For example, ChIP-Seq experiments may display overrepresentation of particular sequences or K-mers (particularity in the case of experiments involving a sequence-specific transcription factors). Likewise BS-Seq experiments have atypical nucleotide distributions, forward reads will be almost devoid of cytosines while reverse reads (in paired-end sequencing) will be largely devoid of guanines.

Pre-alignment Filtering

Raw NGS data is typically provided as fastq or (unaligned) BAM file formats. If necessary, it is possible to convert between these formats using the SAMtools package (Li et al. 2009; www.samtools.sourceforge.net). Prior to alignment data the raw can be filtered to remove sequences corresponding to the adaptors used to generate

the library that may inadvertently be sequenced. This is especially critical in the case of small RNA-seq libraries where it is essential to carry out adaptor trimming due to the small size of the insert. Sequencing accuracy often expresses as base-call quality (confidence that a given base in a read is correctly called) often declines towards the 3′ end of reads. This can be addressed by trimming reads that have low quality base-calls. One approach is to trim on a read-by-read basis where (if needed) the ends of the read are trimmed until only bases with base-calls above the desired threshold remain. Alternatively if it is desired that all read remain the same length (as read length can affect mapping accuracy), one can decide to trim n number of reads from the 3′ end of all reads. Reads sometimes show unusual nucleotide biases in the 5′ ends (possibly due to biases in the ligation efficiency of certain nucleotide sequences to the adaptors) it may be desirable to trim a certain number of bases from the 5′ end as well. These trimming and filtering steps can be carried out using the FASTX-toolkit (www.hannonlab.cshl.edu/fastx_toolkit) which is also available in the Galaxy environment.

Alignment

After filtering the reads are aligned to a reference genome. If available, this should be the same genotype as used in the experiment. Otherwise the most genetically similar genome available should be used. When aligning to a non-identical genome it is generally advantageous to allow for SNPs by specifying a higher number of mismatches to allow for SNPs between the reads and the reference and to use an aligner that allows "gaps" to allow for indels.

A number of alignment programs are available for mapping short reads to a reference genome. These either use a masked hash table strategy such as RMAP (Smith et al. 2008) and MAQ (Li et al. 2008a) or Burrows-wheeler transform such as BWA (Li and Durbin 2009), SOAP (Li et al. 2008b) and Bowtie (Langmead et al. 2009). Other aligners use a hybrid strategy of rough mapping using a hashing based approach then fine map using Smith-Waterman local alignment such as BFAST (Homer et al. 2009) or SMALT (www.sanger.ac.uk/resources/software/smalt). Although many aligners have their own unique output format, it is generally better set the aligner to output the alignment file in BAM or SAM format as these are compatible with most downstream analysis tools.

BS-Seq Alignment

Due to mismatches caused by the conversion process where unmethylated cytosine ends up as thymine, alignment of reads from bisulfite treaded library's reads requires the use of a bisulfite aware mapping strategy. This allows to thymines in the forward reads to map to genomic cytosines and (for paired end libraries) adenine in the

reverse reads to map to genomic guanines. This can be achieved using bisulfite specific alignment packages such as BSMAP (Xi and Li 2009), Bismark (Krueger and Andrews 2011), Methylcoder (Pedersen et al. 2011), or BiSS (Dinh et al. 2012). These packages deal with the problem of cytosine to thymine mismatches using one of two approaches. In the "biased" approach (employed by BSMAP and BiSS) reads are mapped to the reference allowing thymines in the read to map to C in the reference without penalty. In the "unbiased" or "3-letter genome" approach a script converts all the cytosines in the reads and the reference genome to thimines prior to alignment. After alignment a second script restores the cytosines that were originally there. The "unbiased" approach is totally blind to the methylation status and will map reads in methylated and unmethylated loci equally well. On the other hand, the reduced sequence complexity (a 3-letter genome instead of a 4-letter genome) reduces alignment efficiency. The "biased" approach makes full use of the information available and will (all other parameters being equal) align more reads. The drawback of the "biased" approach is that they have more power to uniquely map reads to methylated loci as the cytosines remaining in the reads give more information (as they can only be mapped to a genomic cytosine) than those coming from an unmethylated loci where the cytosines have been converted to thymines (and can thus map to either a cytosine or a thiamine in the genome). For this reason the "biased" approach may overestimate the amount of methylation present.

Methylation Calling

After alignment the methylation level of each cytosine in the genome is determined. This functionality is included with bisulfite alignment packages but can also be done via a custom script or a stand-alone package such as Bis-SNP (which can also call SNPs from BS-Seq data; Liu et al. 2012).

Estimation of Conversion Efficiency

In order for BS-Seq data to be useful, it is imperative that the conversion efficiency of the bisulfite reaction is high enough, otherwise unmethylated regions will appear as methylated in this assay (Lister et al. 2008). The easiest way to estimated conversion efficiency is to calculate the methylation level for cytosines in the chloroplast (as the chloroplast is devoid of DNA methylation). Ideally the average methylation level of cytosines on the chloroplast should be less than 1 % (i.e., corresponding to a conversion rate of over 99 %; Li et al. 2013a).

Downstream Analysis

In plants methylation of CG, CHG, and CHH occurs via distinct pathways and they have (at least partially) distinct distribution and functions. For example, CG methylation is abundant on the gene bodies of highly expressed genes while CHG and

CHH are largely restricted to transposons and other repetitive elements where they appear to have a repressive function (Meyer 2011). Because of this it is often beneficial to split the methylation data by context and to analyze each separately. Analysis can be performed at different levels and in many cases it is useful to look at average methylation over a feature, be it the genome, all genes, all transposons or a particular subset of them. This type of analysis can help to reveal general properties of the biological sample under investigation. An advantage of this type of analysis is that the data are averaged over many loci, so having replicates is not so critical and likewise low coverage data can also be used (Hardcastle 2013).

Differential Methylation Analysis

An alternative approach is to determine regions of the genome where methylation differs between biological samples. Methylation normally occurs at clusters of cytosines in discrete regions of the genome; however within these regions, the exact cytosines methylated will vary from cell to cell and also between biological samples. At the single cytosine level this methylation appears to be largely stochastic but it is correlated with the methylation levels of other cytosines around it (Becker et al. 2011; Schmitz et al. 2011). For this reason instead of looking for differences in methylation of individual cytosines, it is often more biologically informative to define methylated regions that vary between samples (differentially methylated regions, or DMRs). This can be done in different ways. One can either sum or average them methylation level over bins or sliding windows of a defined size or define bins based on genomic features (e.g., genes, Stroud et al. 2013). Other approaches for defining DMRs include defining domains based on a minimal number of methylated cytosines within a certain genomic distance or using hidden Markov or bimodal distribution models (Lister et al. 2008; Schmitz et al. 2011; Li et al. 2013b).

As the distribution of methylation at a particular cytosine is not random, but correlated with the methylation level of adjacent cytosines, it possible to smooth BS-Seq data using a local methylation average (Hansen et al. 2012) This is particularly useful when dealing with low coverage data. Other software packages that may be useful for post alignment analysis of methylation data include methylKit (Akalin et al. 2012) and Repitools (Statham et al. 2010).

ChIP-Seq Analysis

Web Services

As ChIP-Sequencing is quite a widely used technique many of the commonly used tools are implemented in the Galaxy web server environment making at least basic ChIP-seq analysis feasible without resorting to command line tools. Another user-friendly web based platform for ChIP-Seq analysis in plants is PRI-CAT (Muino et al. 2011).

Post-Processing

After the mapping of reads to genome those mapping to multiple genomic loci should be discarded, using, for example, SAMtools (Li et al. 2009). The resulting SAM or BAM file should then be converted to a BED file that is required by most peak calling programs using, for example, BEDtools program (Quinlan and Hall 2010).

Peak Calling

The next step is to determine the regions of the genome enriched for the precipitated antigen using a peak calling program. Most peak callers require a sample file (reads from an immunoprecipitated library) and if available, a control file (reads from input or other control library). A wide variety of peak calling algorithms are available (Wilbanks and Facciotti 2010). MACS (Zhang et al. 2008) is widely used for calling transcription factor peaks and is available in the Galaxy environment. One complication in ChIP-Seq analysis is the presence of multiple identical reads. These can occur for two reasons, firstly they may be PCR duplicates, originating from the same fragment of genomic DNA (and should ideally be filtered out prior to peak calling) or they may simply reflect high levels of enrichment of the antigen at that particular locus (and as such represent real signal). While removal of duplicates avoids the chances of artifacts due to the PCR amplification, the signal to noise ratio will be reduced if too many reads originating from independent DNA molecules are removed (Muino et al. 2011). The best approach will depend to some extent on the properties of the individual ChIP-Seq library, but one compromise is to set a threshold for number of duplicates allowed and only discard duplicates exceeding it.

Compared to transcription factors, some histone modifications such as H3K36me2 and H3K27me3 occur in long, diffuse domains that are poorly detected by some peak-calling algorithms. This appears to be because most peak callers were designed to detect sharp peaks typical of transcription factors. To address this several peak callers have been specifically developed to detect enrichment of the broad domains typical of histone modifications. These include SICER (Zang et al. 2009, also available in Galaxy) which uses a sliding window approach and RSEG (Song and Smith 2011) and BCP (Xing et al. 2012) which use hidden Markov modeling and Bayesian change point approaches.

The depth of sequencing required to identify all the peaks of interest in a ChIP-Seq experiment depends on a number of factors, including the size of genome, the efficiency of the immunoprecipitation step (i.e., signal to noise ratio), and what the target antigen was. Although it is hard to determine exactly where the saturation point is, a reasonable estimate can be obtained by determining the total number of peaks called from the entire library. This is then repeating by calling peaks using progressively smaller fractions of the library. The number of peaks called vs library size can then be plotted. As library size increases the line representing the number of peaks call will never completely plateau, but it should flatten out when a reasonable depth of sequencing is obtained (Liu et al. 2010).

Higher-Level Analysis

It is often useful to identify peaks with differential enrichment between biological samples. Some peak callers include this functionality, e.g. SICER (Zang et al. 2009); otherwise, it can be estimated by various packages such as ChIPDiff (Xu et al. 2008) and EdgeR (Robinson et al. 2010). Further analysis can involve determining which genes (or other genomic features) are associated with the peaks identified using the ChIPpeakAnno package or similar (Zhu et al. 2010). The Repitools package (Statham et al. 2010) is also useful for identifying properties of interest in ChIP-seq data such as distribution of peaks relative to transcription start sites or other features of interest.

Nucleosome Positioning Analysis

MNase-seq data has properties substantially similar to ChIP-seq (with the exception that paired end data is usually gathered to improve the accuracy of nucleosome positioning). For this reason the same data processing steps can be performed as described in the sections "Web Services" and "Post-Processing." Although standard peak-callers such as SICER can be used to infere the positions of the nucleosomes, a number of specialized MNase-Seq tools are available such as nucleR (Flores and Orozco 2011) and PING (Woo et al. 2013) which can estimate nucleosome occupancy at different sites and how well positioned nucleosomes are. Another option is the Nseq program which provides users with a GUI interface (Nellore et al. 2012). As with ChIP-Seq, plotting the number of nucleosomes called vs number of reads library can be used to infer the depth of sequencing required.

Small RNA-Seq Analyses

Due to their small size it is critical that small RNA-Seq reads are trimmed to remove adapter sequences. Following this the reads should be filtered based on their length to retain only those of length corresponding to small RNAs (e.g., 18–34 nt). Both these steps can be done using the FASTX toolkit or similar. Small RNAs generally originate from repetitive regions and this, along with their small size, makes it difficult to determine from where in the genome they originate. For this reason small RNAs are generally aligned to the genome using an aligner such as BWA or Bowtie without allowing for mismatches but allowing for reads mapping to multiple locations in the genome. Differences in enrichment of between biological samples can be estimated using packages such as edgeR (McCarthy et al. 2012) or baySeq (Hardcastle and Kelly 2013).

Data Visualization

Most analysis tools allow processed data to be exported as WIG or BED format files which can be displayed together with annotations of interest (i.e., genes, expression data, etc.) in a genome browser. These files can be uploaded to a web based genome browser such as gbrowse implemented by TAIR for *Arabidopsis* (www.arabidopsis.org) or the EPIC-COGE browser, which supports a wide range of species (http://genomevolution.org/r/939v). Alternatively data can be visualized using software such as Integrated Genomics Viewer (IGV) (Robinson et al. 2011) or Integrated genome Browser (Nicol et al. 2009).

References

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A et al (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13, R87

Autran D, Baroux C, Raissig MT, Lenormand T, Wittig M, Grob S et al (2011) Maternal epigenetic pathways control parental contributions to *Arabidopsis* early embryogenesis. Cell 145:707–719

Ballestar E, Paz MF, Valle L, Wei S, Fraga MF, Espada J et al (2003) Methyl-CpG binding proteins identify novel sites of epigenetic inactivation in human cancer. EMBO J 22:6335–6345

Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. Cell Res 21:381–395

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z et al (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837

Becker C, Hagmann J, Muller J, Koenig D, Stegle O, Borgwardt K et al (2011) Spontaneous epigenetic variation in the *Arabidopsis* thaliana methylome. Nature 480:245–249

Benhamed M, Martin-Magniette ML, Taconnat L, Bitton F, Servet C, De Clercq R et al (2008) Genome-scale *Arabidopsis* promoter array identifies targets of the histone acetyltransferase GCN5. Plant J 56:493–504

Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16:545–552

Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, Liu JS et al (2002) Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci U S A 99:8695–8700

Bird AP (1984) DNA methylation–how important in gene control? Nature 307:503–504

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M et al (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol Chapter 19, Unit 19 0 1–21

Boccara M, Sarazin A, Billoud B, Jolly V, Martienssen R, Baulcombe D et al (2007) New approaches for the analysis of *Arabidopsis* thaliana small RNAs. Biochimie 89:1252–1256

Bossdorf O, Richards CL, Pigliucci M (2008) Epigenetics for ecologists. Ecol Lett 11:106–115

Calarco JP, Borges F, Donoghue MT, Van Ex F, Jullien PE, Lopes T et al (2012) Reprogramming of DNA methylation in pollen guides epigenetic inheritance via small RNA. Cell 151:194–205

Chen YR, Zheng Y, Liu B, Zhong S, Giovannoni J, Fei Z (2012) A cost-effective method for Illumina small RNA-Seq library preparation using T4 RNA ligase 1 adenylated adapters. Plant Methods 8:41

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y et al (2010) Relationship between nucleosome positioning and DNA methylation. Nature 466:388–392

Clark SJ, Harrison J, Paul CL, Frommer M (1994) High sensitivity mapping of methylated cytosines. Nucleic Acids Res 22:2990–2997

Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD et al (2008) Shotgun bisulphite sequencing of the *Arabidopsis* genome reveals DNA methylation patterning. Nature 452:215–219

Cortijo S, Wardenaar R, Colomé-Tatché M, Johannes F, and Colot V (2014) Genome-wide analysis of DNA methylation in Arabidopsis using MeDIP-chip. Methods Mol Biol. 1112:125–149

Costa S, Shaw P (2007) 'Open minded' cells: how cells can change fate. Trends Cell Biol 17:101–106

Cubas P, Vincent C, Coen E (1999) An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401:157–161

Dinh HQ, Dubin M, Sedlazeck FJ, Lettner N, Mittelsten Scheid O, von Haeseler A (2012) Advanced methylome analysis after bisulfite deep sequencing: an example in *Arabidopsis*. PLoS One 7:e41528

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR et al (2012) Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci U S A 109:E2183–E2191

Eichten SR, Swanson-Wagner RA, Schnable JC, Waters AJ, Hermanson PJ, Liu S et al (2011) Heritable epigenetic variation among maize inbreds. PLoS Genet 7:e1002372

English JJ, Mueller E, Baulcombe DC (1996) Suppression of virus accumulation in transgenic plants exhibiting silencing of nuclear genes. Plant Cell 8:179–188

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG et al (2010) Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 107:8689–8694

Flores O, Orozco M (2011) nucleR: a package for non-parametric nucleosome positioning. Bioinformatics 27:2149–2150

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA et al (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7:461–465

Gilfillan GD, Hughes T, Sheng Y, Hjorthaug HS, Straub T, Gervin K et al (2012) Limitations and possibilities of low cell number ChIP-seq. BMC Genomics 13:645

Gitan RS, Shi H, Chen CM, Yan PS, Huang TH (2002) Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. Genome Res 12:158–164

Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C et al (2008) Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. Methods 44:3–12

Hansen KD, Langmead B, Irizarry RA (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 13:R83

Hardcastle TJ (2013) High-throughput sequencing of cytosine methylation in plant DNA. Plant Methods 9:16

Hardcastle TJ, Kelly KA (2013) Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution. BMC Bioinformatics 14:135

Haring M, Offermann S, Danker T, Horst I, Peterhansel C, Stam M (2007) Chromatin immunoprecipitation: optimization, quantitative analysis and data normalization. Plant Methods 3:11

He G, Zhu X, Elling AA, Chen L, Wang X, Guo L et al (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. Plant Cell 22:17–33

Ho JW, Bishop E, Karchenko PV, Negre N, White KP, Park PJ (2011) ChIP-chip versus ChIP-seq: lessons for experimental design and data analysis. BMC Genomics 12:134

Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4:e7767

Hsieh TF, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL et al (2009) Genome-wide demethylation of *Arabidopsis* endosperm. Science 324:1451–1454

Ibarra CA, Feng X, Schoft VK, Hsieh TF, Uzawa R, Rodrigues JA et al (2012) Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. Science 337:1360–1364

Jacobsen SE, Meyerowitz EM (1997) Hypermethylated SUPERMAN epigenetic alleles in *Arabidopsis*. Science 277:1100–1103

Jenuwein T, Allis CD (2001) Translating the histone code. Science 293:1074–1080

Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10:161–172

Kalisz S, Kramer EM (2008) Variation and constraint in plant evolution and development. Heredity (Edinb) 100:171–177

Kanno T, Huettel B, Mette MF, Aufsatz W, Jaligot E, Daxinger L et al (2005) Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nat Genet 37:761–765

Kaufmann K, Muino JM, Osteras M, Farinelli L, Krajewski P, Angenent GC (2010) Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nat Protoc 5:457–472

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27:1571–1572

Krueger F, Kreck B, Franke A, Andrews SR (2012) DNA methylome analysis using short bisulfite sequencing data. Nat Methods 9:145–151

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760

Li H, Ruan J, Durbin R (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858

Li R, Li Y, Kristiansen K, Wang J (2008b) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714

Li X, Wang X, He K, Ma Y, Su N, He H et al (2008c) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. Plant Cell 20:259–276

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/Map format and SAMtools. Bioinformatics 25:2078–2079

Li P, Demirci F, Mahalingam G, Demirci C, Nakano M, Meyers BC (2013a) An integrated workflow for DNA methylation analysis. J Genet Genomics 40:249–260

Li S, Garrett-Bakelman FE, Akalin A, Zumbo P, Levine R, To BL et al (2013b) An optimized algorithm for detecting and annotating regional differential methylation. BMC Bioinformatics 14 Suppl 5, S10

Lindroth AM, Cao X, Jackson JP, Zilberman D, McCallum CM, Henikoff S et al (2001) Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science 292:2077–2080

Linsen SE, Cuppen E (2012) Methods for small RNA preparation for digital gene expression profiling by next-generation sequencing. Methods Mol Biol 822:205–217

Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, McCombie WR et al (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430:471–476

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH et al (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell 133:523–536

Liu ET, Pott S, Huss M (2010) Q&A: ChIP-seq technologies and the study of gene regulation. BMC Biol 8:56

Liu Y, Siegmund KD, Laird PW, Berman BP (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol 13:R61

Lu C, Meyers BC, Green PJ (2007) Construction of small RNA cDNA libraries for deep sequencing. Methods 43:110–117

Luo C, Sidote DJ, Zhang Y, Kerstetter RA, Michael TP, Lam E (2012) Integrative analysis of chromatin states in *Arabidopsis* identified potential regulatory mechanisms for natural antisense transcript production. Plant J 73:77–90

Madlung A, Comai L (2004) The effect of stress on genome regulation and structure. Ann Bot 94:481–495

Martienssen RA, Colot V (2001) DNA methylation and epigenetic inheritance in plants and filamentous fungi. Science 293:1070–1074

Martienssen RA, Doerge RW, Colot V (2005) Epigenomic mapping in *Arabidopsis* using tiling microarrays. Chromosome Res 13:299–308

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke AJ (2009) RNA-mediated chromatin-based silencing in plants. Curr Opin Cell Biol 21:367–376

McCarthy DJ, Chen Y, Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 40:4288–4297

McClintock B (1950) The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A 36:344–355

McClintock B (1956) Controlling elements and the gene. Cold Spring Harb Symp Quant Biol 21:197–216

Meyer P (2011) DNA methylation systems and targets in plants. FEBS Lett 585:2008–2015

Moghaddam AM, Roudier F, Seifert M, Berard C, Magniette ML, Ashtiyani RK et al (2011) Additive inheritance of histone modifications in *Arabidopsis* thaliana intra-specific hybrids. Plant J 67:691–700

Muino JM, Hoogstraat M, van Ham RC, van Dijk AD (2011) PRI-CAT: a web-tool for the analysis, storage and visualization of plant ChIP-seq experiments. Nucleic Acids Res 39:W524–W527

Nellore A, Bobkov K, Howe E, Pankov A, Diaz A, Song JS (2012) NSeq: a multithreaded Java application for finding positioned nucleosomes from sequencing data. Front Genet 3:320

Nicol JW, Helt GA, Blanchard SG Jr, Raja A, Loraine AE (2009) The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 25:2730–2731

Oh E, Kang H, Yamaguchi S, Park J, Lee D, Kamiya Y et al (2009) Genome-wide analysis of genes targeted by PHYTOCHROME INTERACTING FACTOR 3-LIKE5 during seed germination in *Arabidopsis*. Plant Cell 21:403–419

O'Neill LP, Turner BM (2003) Immunoprecipitation of native chromatin: NChIP. Methods 31:76–82

Parent JS, Martinez de Alba AE, Vaucheret H (2012) The origin and effect of small RNA signaling in plants. Front Plant Sci 3:179

Paszkowski J, Grossniklaus U (2011) Selected aspects of transgenerational epigenetic inheritance and resetting in plants. Curr Opin Plant Biol 14:195–203

Pedersen B, Hsieh TF, Ibarra C, Fischer RL (2011) MethylCoder: software pipeline for bisulfite-treated sequences. Bioinformatics 27:2435–2436

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842

Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J (2011) A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470:279–283

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G et al (2011) Integrative genomics viewer. Nat Biotechnol 29:24–26

Rodrigues JA, Ruan R, Nishimura T, Sharma MK, Sharma R, Ronald PC et al (2013) Imprinted expression of genes and small RNA is associated with localized hypomethylation of the maternal genome in rice endosperm. Proc Natl Acad Sci U S A 110:7934–7939

Ronemus MJ, Galbiati M, Ticknor C, Chen J, Dellaporta SL (1996) Demethylation-induced developmental pleiotropy in *Arabidopsis*. Science 273:654–657

Roudier F, Ahmed I, Berard C, Sarazin A, Mary-Huard T, Cortijo S et al (2011) Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. EMBO J 30:1928–1938

Sani E, Herzyk P, Perrella G, Colot V, Amtmann A (2013) Hyperosmotic priming of *Arabidopsis* seedlings establishes a long-term somatic memory accompanied by specific changes of the epigenome. Genome Biol 14:R59

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O et al (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science 334:369–373

Schmitz RJ, He Y, Valdes-Lopez O, Khan SM, Joshi T, Urich MA et al (2013a) Epigenome-wide inheritance of cytosine methylation variants in a recombinant inbred population. Genome Res

Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O et al (2013b) Patterns of population epigenomic diversity. Nature 495:193–198

Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, Yau P et al (2006) Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Res 34:528–542

Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128

Song Q, Smith AD (2011) Identifying dispersed epigenomic domains from ChIP-Seq data. Bioinformatics 27:870–871

Statham AL, Strbenac D, Coolen MW, Stirzaker C, Clark SJ, Robinson MD (2010) Repitools: an R package for the analysis of enrichment-based epigenomic data. Bioinformatics 26:1662–1663

Stroud H, Greenberg MV, Feng S, Bernatavichute YV, Jacobsen SE (2013) Comprehensive analysis of silencing mutants reveals complex regulation of the *Arabidopsis* methylome. Cell 152:352–364

Sung S, Amasino RM (2004) Vernalization and epigenetics: how plants remember winter. Curr Opin Plant Biol 7:4–10

Takeda S, Tadele Z, Hofmann I, Probst AV, Angelis KJ, Kaya H et al (2004) BRU1, a novel link between responses to DNA damage and epigenetic gene silencing in *Arabidopsis*. Genes Dev 18:782–793

Takuno S, Gaut BS (2013) Gene body methylation is conserved between plant orthologs and is of evolutionary consequence. Proc Natl Acad Sci U S A 110:1797–1802

Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat Struct Mol Biol 14:1025–1040

Teixeira FK, Heredia F, Sarazin A, Roudier F, Boccara M, Ciaudo C et al (2009) A role for RNAi in the selective correction of DNA methylation defects. Science 323:1600–1604

Tompa R, McCallum CM, Delrow J, Henikoff JG, van Steensel B, Henikoff S (2002) Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. Curr Biol 12:65–68

Turck F, Roudier F, Farrona S, Martin-Magniette ML, Guillaume E, Buisine N et al (2007) *Arabidopsis* TFL2/LHP1 specifically associates with genes marked by trimethylation of histone H3 lysine 27. PLoS Genet 3:e86

Vaughn MW, Tanurdzic M, Lippman Z, Jiang H, Carrasquillo R, Rabinowicz PD et al (2007) Epigenetic natural variation in *Arabidopsis* thaliana. PLoS Biol 5:e174

Vu TM, Nakamura M, Calarco JP, Susaki D, Lim PQ, Kinoshita T et al (2013) RNA-directed DNA methylation regulates parental genomic imprinting at several loci in *Arabidopsis*. Development 140:2953–2960

Wassenegger M, Heimes S, Riedel L, Sanger HL (1994) RNA-directed de novo methylation of genomic sequences in plants. Cell 76:567–576

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL et al (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37:853–862

Wei G, Hu G, Cui K, Zhao K (2012) Genome-wide mapping of nucleosome occupancy, histone modifications, and gene expression using next-generation sequencing technology. Methods Enzymol 513:297–313

Weinmann AS, Yan PS, Oberley MJ, Huang TH, Farnham PJ (2002) Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. Genes Dev 16:235–244

Wilbanks EG, Facciotti MT (2010) Evaluation of algorithm performance in ChIP-seq peak detection. PLoS One 5:e11471

Wollmann H, Holec S, Alden K, Clarke ND, Jacques PE, Berger F (2012) Dynamic deposition of histone variant H3.3 accompanies developmental remodeling of the *Arabidopsis* transcriptome. PLoS Genet 8:e1002658

Woo S, Zhang X, Sauteraud R, Robert F, Gottardo R (2013) PING 2.0: an R/Bioconductor package for nucleosome positioning using next-generation sequencing data. Bioinformatics 29(16):2049–2050

Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10:232

Xing H, Mo Y, Liao W, Zhang MQ (2012) Genome-wide localization of protein-DNA binding and histone modification by a Bayesian change-point method with ChIP-seq data. PLoS Comput Biol 8:e1002613

Xu H, Wei CL, Lin F, Sung WK (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics 24:2344–2349

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25:1952–1958

Zemach A, Kim MY, Hsieh PH, Coleman-Derr D, Eshed-Williams L, Thao K et al (2013) The *Arabidopsis* nucleosome remodeler DDM1 allows DNA methyltransferases to access H1-containing heterochromatin. Cell 153:193–205

Zeng PY, Vakoc CR, Chen ZC, Blobel GA, Berger SL (2006) In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. Biotechniques 41:694, 696, 698

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137

Zhang YY, Fischer M, Colot V, Bossdorf O (2013) Epigenetic variation creates potential for evolution of plant phenotypic plasticity. New Phytol 197:314–322

Zheng Y, Ren N, Wang H, Stromberg AJ, Perry SE (2009) Global identification of targets of the *Arabidopsis* MADS domain protein AGAMOUS-Like15. Plant Cell 21:2563–2577

Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS et al (2010) ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 11:237

Zhu JY, Sun Y, Wang ZY (2012) Genome-wide identification of transcription factor-binding sites in plants using chromatin immunoprecipitation followed by microarray (ChIP-chip) or sequencing (ChIP-seq). Methods Mol Biol 876:173–188

Chapter 3
Whole Genome Sequencing to Identify Genes and QTL in Rice

Ryohei Terauchi, Akira Abe, Hiroki Takagi, Muluneh Tamiru, Rym Fekih, Satoshi Natsume, Hiroki Yaegashi, Shunichi Kosugi, Hiroyuki Kanzaki, Hideo Matsumura, Hiromasa Saitoh, Kentaro Yoshida, Liliana Cano, and Sophien Kamoun

Overview of Genetic Analysis for Identifying Genes

One of the major purposes of genetic analysis is to infer an alteration in the genetic material that is responsible for the phenotypic changes observed in an organism under study. This has been routinely addressed by genetic association studies. For example, let's assume that we have a population of individuals segregating in two phenotypic variants, and that our interest is to identify the gene responsible for this phenotypic difference. For this purpose, we first divide the population into two phenotypic groups and then look for the genetic variation that shows statistically significant association with the groups. Since genes are arranged linearly on chromosomes, two loci that are physically close to each other are more likely inherited together, whereas distantly located loci tend to be inherited independently due to recombination occurring between the two loci (Bateson et al. 1905; Morgan 1910; Lobo and Shaw 2008). Therefore, once an association is identified between a genetic variation and the phenotype under investigation, we infer that physical location of the causative gene controlling the phenotype is close (linked) to the identified genetic variation. A common practice is to first identify as many genetic variations

R. Terauchi, Ph.D. (✉) • A. Abe, Ph.D. • H. Takagi, Ph.D. • M. Tamiru, Ph.D.
R. Fekih, Ph.D. • S. Natsume • H. Yaegashi • H. Kanzaki, Ph.D. • H. Saitoh, Ph.D.
Division of Genomics and Breeding, Iwate Biotechnology Research Center,
Narita 22-174-4, Kitakami 024-0003, Iwate, Japan
e-mail: terauchi@ibrc.or.jp

S. Kosugi, Ph.D.
Kazusa DNA Research Institute, Kisarazu, Chiba, Japan

H. Matsumura, Ph.D.
Gene Research Center, Shinshu University, Ueda, Nagano, Japan

K. Yoshida, Ph.D. • L. Cano, Ph.D. • S. Kamoun, Ph.D.
The Sainsbury Laboratory, Norwich Research Park, Norwich, Norfork, UK

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_3

segregating among the individuals of the study, and use these variations as "genetic markers" to test their association with the phenotype. Following identification of genetic markers that show association with a phenotype, we explore their vicinity to identify the very genetic change that is responsible for the phenotypic variation.

Two major approaches have been largely employed in genetic association studies. The first is applied to progeny derived from a cross between known parents; therefore, it is most widely used for gene isolation from crop species that are amenable to artificial crossing. Typically, crossing of two inbred parental lines results in F1, which is self-fertilized to generate F2 progeny. Using a large number of F2 progeny segregating for a particular phenotype, the association between the phenotype and genetic markers is examined. This approach addresses linkage (co-segregation) of phenotypes and markers from the parents to progeny, thus is usually called "linkage study." The second genetic association approach does not involve crossing, and is applied to a population of individuals with unknown relationships to each other. This approach is commonly called "association study," and whole genome association study (WGAS) has been widely used for gene identification in humans and other organisms. In WGAS, the population is divided into "case" and "control" groups to reveal markers that are associated with the "case"/"control" dichotomy. Each approach has its advantages and disadvantages. Linkage analysis is usually carried out over two generations (parents–offspring). As a result, the number of recombination occurring between the two generations is limited. In contrast, association analysis depends on individuals whose common ancestor traces back a large number of generations, ensuring that the number of recombinations among the individuals under study is large. The difference in the number of recombinations affects performance of the analysis. Owing to a higher level of linkage disequilibrium (LD), linkage analysis can be powerful in finding an approximate position of a causative gene using a small number of markers, but requires additional efforts in identifying the causative gene itself. On the other hand, low levels of LD make association analysis less suited for inferring the approximate position of causative genes, but it is more powerful in identifying the causative gene, provided that a large number of markers are available. Consequently, the combination of linkage analysis and association study has proved powerful as the two approaches complement each other.

Genetic Markers to Become Obsolete?

For linkage analysis and association study, the availability of a large number of genetic markers is a prerequisite for successful analysis. As a result, researchers in the field of genetics have devoted considerable amount of time and resources over the years to develop such markers (Avise 1994). Development of DNA technology in the 1970s enabled the use of Restriction Fragment Length Polymorphisms (RFLP) markers. This was followed by the invention of Polymerase Chain Reaction (PCR) and discovery of ubiquitous distribution of di- tri-nucleotide simple sequence repeats (SSR) in eukaryotic genomes, which allowed the application of highly

variable markers called SSR or microsatellite markers. Development and advances in automated DNA sequencing technology enabled the identification of unlimited numbers of single nucleotide polymorphisms (SNPs) in the genome that can be used as markers called SNP markers. Continuous efforts have been made to generate large number of genetic markers covering the entire genome, while at the same time trying to make their scoring easier and cheaper. Nevertheless, the available genetic markers still represent a small proportion of the entire genetic code of an organism, forcing researchers to make inferences about properties of a whole genome based on the markers that are basically a limited number of sample points selected from the genome. However, thanks to the recent development of WGS technology, this situation is beginning to change, and would hopefully free researchers from their dependence on classical genetic markers.

Rice Genetic Resources at IBRC

In order to fully exploit genomics approaches for efficient crop improvement, the availability of suitable biological materials is essential. Genetic materials representing a wide genetic variation of the species under study should be carefully generated, maintained, and scored for their phenotypes. At the Iwate Biotechnology Research Center (IBRC), we have been generating and maintaining two sets of rice genetic resources to accelerate improvement of elite rice cultivars. The first set of materials include ethylmethanesulfonate (EMS)-induced mutant lines (Rakshit et al. 2010). We treated immature embryo of flowers of a Northern Japan elite rice (*Oryza sativa* spp. *japonica*) cultivar "Hitomebore" with 0.75 % EMS. The matured seeds were planted to generate M1 individuals, which were self-fertilized to obtain M2 seeds. The M2 plants were further self-fertilized to obtain M3 and subsequent generations, and we are currently maintaining seeds of over 12,000 mutant lines at M3–M5 generations. These mutant lines show a wide range of phenotypic diversity, particularly of traits of agronomic importance. The second set of materials represent recombinant inbred lines (RILs) obtained by crossing of "Hitomebore" to 22 rice cultivars representing a wide genetic variation of *O. sativa*. We currently have a total of 3,172 RILs at the F5 to F7 generations. These resources are being used for isolation of important genes and QTL, as well as to develop WGS-based methods for accelerating crop breeding.

MutMap

Using the EMS-mutant lines of "Hitomebore" rice cultivar, we set out to rapidly identify the causal mutation responsible for a given mutant phenotypic trait of agronomic importance. For this purpose, we developed the MutMap method (Abe et al. 2012). In MutMap, a mutant of interest is crossed to the parental line used for

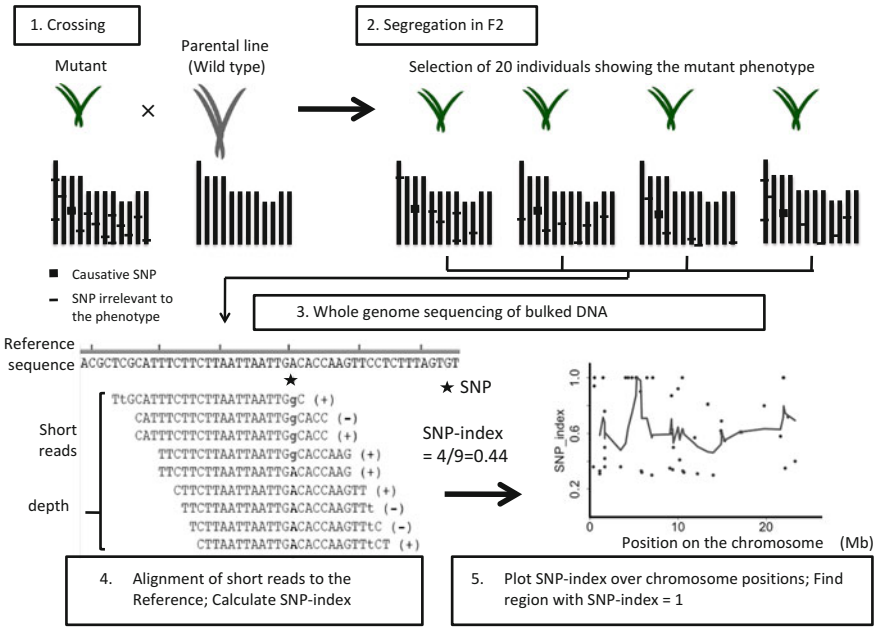

Fig. 3.1 A simplified scheme of MutMap. *1*. A mutant showing a phenotype of interest is crossed to the parental line used for mutagenesis to obtain F2. Here we take a dwarf mutant caused by recessive mutation as an example. The mutant chromosomes with the mutations incorporated by mutagenesis and wild type individual chromosomes are shown. *2*. Among the F2 progeny segregating for wild type and mutant phenotypes, we focus on the mutant F2 progeny. Mutant F2 individuals inherit the causative mutation in homozygous state whereas mutations that are irrelevant to the phenotype are inherited in 1:1 ratio. *3*. DNA of >20 mutant F2 progeny are bulked and sequenced by Illumina sequencer. *4*. The resulting short reads are aligned to the reference sequence of wild type parental line, and SNP-index is calculated. *5*. SNP-index plots are generated to visualize the relationship between SNP-index and chromosome position. Sliding window analysis is applied to see the average value of SNP-index in a given genome interval. A peak of SNP-index suggests the position of causative mutation responsible for the mutant phenotype

mutagenesis (Fig. 3.1). If the phenotype is caused by a single recessive mutation, the F2 progeny segregates 3:1 (wild type to mutant phenotypes). We extract DNA from >20 mutant F2 individuals, mix the DNA in equal proportion to make a DNA-bulk, which is subsequently sequenced by Illumina DNA sequencer to a depth of at least 10× coverage of the genome. Since the rice genome size is about 380 Mb, we usually generate about 5 Gb short reads for each DNA-bulk. The resulting short reads of 76–100 bp in size are aligned to the reference sequence of the parental line "Hitomebore."

To facilitate the identification of the causative mutation, we introduced the concept of "SNP-index," which is the ratio of short reads containing SNPs to the total reads covering a particular position of the genome. If all short reads covering a particular genomic position have an identical sequence to the reference, the SNP-index is 0.

By contrast, if all the short reads have an SNP different from the reference sequence, the SNP-index is 1. Since the causative SNP responsible for the mutant phenotype should be inherited by all the F2 mutant progeny in homozygous state, short reads of bulked DNA corresponding to such an SNP should have SNP-index = 1, whereas SNPs not relevant to the phenotype should segregate 1:1 among the progeny, resulting in SNP-index = 0.5. The genomic region(s) tightly linked to the causative SNP are dragged with the causative mutation, thus the SNPs residing in such region should have SNP-index >0.5. Thus, if we graphically plot the relationships between SNP-index and chromosomal positions, we would observe a peak of SNP-index that is clustered around the position of the causative SNP. After locating the SNPs with SNP-index = 1, we scrutinize the genes harboring the SNPs, and identify the most likely candidate. Since WGS allows us to follow the inheritance of all the SNPs incorporated by mutagenesis, MutMap analysis does not require any genetic markers. Linkage of SNPs will allow us to visually identify the SNP-index peaks on a graph.

Practically MutMap requires only a total of around 100 F2 progeny to score the phenotype and a single WGS for identification of the causative mutation. Therefore, MutMap circumvents the time-consuming and laborious steps of conventional marker-based linkage analysis. Furthermore, MutMap is applied to F2 derived from the cross of a mutant of interest to the parental line used for mutagenesis. This procedure enables the identification of mutations that cause subtle quantitative changes of phenotype relevant to agronomic traits. This feature makes MutMap a more efficient method to isolate causative mutations with quantitative effects than SHOREmap (Shneeberger et al. 2009), another bulked DNA sequencing method, in which mutant lines are crossed to a distantly related line.

MutMap+

As discussed above, MutMap application requires crossing to the wild type parental line. To address early death or sterile mutants that don't allow crossing, we recently developed MutMap+ (Fekih et al. 2013) as an extension of MutMap. MutMap+ involves identification of a causal mutation by comparing SNP-index plots of two DNA-bulks, mutant bulk and wild type bulk, obtained from a segregating M3 progeny that is derived from a self-fertilized heterozygous M2 individual. MutMap+ analysis starts by the identification of a phenotype of interest in a small number of segregating M2 individuals (about 10) obtained by selfing of each M1 line. Then, the wild type siblings of the identified mutants are left to self-fertilize and grow to maturity to generate M3 seeds. If the mutation is caused by a single recessive gene, two-third of the wild type M2 individuals are expected to harbor the causal mutation in heterozygous state. The M3 progeny established from seeds of these heterozygous individuals segregate 3:1 for wild type and mutant phenotypes. Accordingly, about 100 M3 individuals are grown separately for each M2 line, and for those that segregate for the phenotype of interest, we make two sets of DNA-bulks: one from ~20 mutant M3 progeny and the other from ~20 wild type M3 progeny, both of

which are derived from a single heterozygous M2 plant. The two DNA-bulks are separately sequenced, aligned to the reference, and SNP-index graphs are plotted as in MutMap analysis and compared to each other. A genomic region showing different patterns of SNP-index plots between the two bulks points to the location of the causal mutation differentiating the mutant from the wild type.

For each M2 individual, half of the mutations incorporated by mutagenesis are randomly fixed to homozygote state. Therefore, selfing of an M2 individual results in a large chromosome regions exhibiting SNP-index = 1 in M3 generation, making it difficult to locate the position of causative mutation. However, genomic regions exhibiting SNP-index = 1 by random fixation should be shared by all the M3 progeny, whereas the region showing SNP-index = 1 caused by bulking of mutant progeny is specific to the mutant DNA-bulk. We can thus identify location of the causative mutation by comparing SNP-index plots of the mutant and wild type bulks of M3 progeny. MutMap+ does not require crossing, so it is particularly useful for identifying mutations that cause early stage lethality and infertility thereby hampering crossing. This method is also applicable to crops that are not amenable to artificial crossing. MutMap+ is also potentially useful to isolate dominant mutations. If the mutation is dominant, the expected SNP-index value for mutant type M3 is 0.66 whereas that for the wild type is 0.

MutMap-Gap

To apply MutMap and MutMap+ for the identification of candidate genes, we need a reference sequence of the parental line. In most cases, the cultivars used for mutagenesis are not the same as the ones for which an accurate reference genome is publicly available. In rice, a highly accurate genome sequence is available for a cultivar "Nipponbare" (International Rice Genome Sequencing Project 2005), but not for the cultivar "Hitomebore" used in our studies. Therefore, we generated a pseudo-reference sequence of the cultivar "Hitomebore," by first obtaining short sequence reads of a "Hitomebore" wild type plant and aligning them to the reference genome of "Nipponbare." After identifying all the SNPs between "Hitomebore" and "Nipponbare," the nucleotides of "Nipponbare" were replaced by those of "Hitomebore" at all the sites of SNPs to make the "Hitomebore" reference sequence. Consequently, this "Hitomebore" reference sequence is useful to identify mutations residing in the genomic regions that are conserved between "Hitomebore" and "Nipponbare." However, if the mutation of interest resides in the genomic region present in "Hitomebore" but absent from "Nipponbare," we cannot identify such mutations by using this pseudo-reference sequence. To solve this problem, we developed a method we named MutMap-Gap, which combines MutMap and local de novo assembly (Takagi et al. 2013a).

For MutMap-Gap analysis, we first apply MutMap to a "Hitomebore" mutant of interest, obtain an SNP-index plots, and identify a peak of SNP-index corresponding to the possible genomic region of the causative mutation. If after scrutiny of

candidate SNPs around the SNP-index peak region we cannot identify any promising SNPs in the "Hitomebore" pseudo-reference, this should prompt us to suspect that the causative mutation may reside in "Hitomebore"-specific region linked to the region identified by MutMap. To explore this possibility, we can apply local de novo assembly by using (1) short reads of wild type "Hitomebore" mapped to the genomic region corresponding to the SNP-index peak as well as (2) short reads of "Hitomebore" that could not be mapped (unmapped) to the "Nipponbare" reference. These unmapped reads are likely to be derived from "Hitomobore"-specific genomic region absent from "Nipponbare." After contigs are generated by the local de novo assembly, short reads of bulked DNA of mutant plants are aligned against the newly prepared reference sequence comprising "Hitomebore" pseudo-reference plus the newly generated contigs. This MutMap-Gap analysis may locate a contig(s) not present in "Nipponbare" that harbors an SNP(s) with SNP-index = 1. By applying MutMap-Gap to "Hitomebore," we successfully identified an SNP in the resistance (R-) gene *Pii* that confers resistance against rice blast fungus with *AVR-Pii*. The complete Hitomebore *Pii* gene region was not represented in "Nipponbare" genome, and it was recovered only by using MutMap-Gap. MutMap-Gap is particularly useful to identify mutations in highly variable genomic regions like *R*-gene clusters that are known to be rapidly evolving.

QTL-Seq

The majority of agronomically important traits are controlled by multiple genes called quantitative trait loci (QTL) (Falconer and Mackay 1996) each with a relatively minor effect. Identification of QTL is an important task in plant breeding, and has been carried out mainly by linkage analysis. Following a cross of two distantly related varieties, F2 or RILs are generated, and their phenotype scored. Using genetic markers, association between trait values and marker genotypes are studied. Due to the necessity of large number of genetic markers, the two mapping parents are usually selected from genetically distantly related lines. However, such parents tend to have differences in multiple QTL, making isolation of individual QTL difficult.

QTL-seq (Takagi et al. 2013b) is a WGS-based method of QTL identification based on bulked-segregant analysis (Giovannoni et al. 1991; Michelmore et al. 1991; Mansur et al. 1993; Darvasi and Soller 1994). First, we cross two cultivars with different trait values and obtain the progeny of F2 generation or RILs (Fig. 3.2). Trait values are measured in the progeny. If the trait is controlled by multiple QTL, frequency distribution of trait values will be close to Normal (Gaussian) distribution. Here we focus on the individual lines that belong to the upper and lower tails of the distribution. We then bulk the DNA of the individuals belonging to the upper tail to make High bulk DNA. Similarly we bulk the DNA of the lower tail individuals to make Low bulk DNA. DNA extracted from High and Low bulks are separately subjected to WGS, and the resulting short reads are aligned to the reference sequence of either of the parental lines (e.g., Cultivar A). The SNP-index as defined

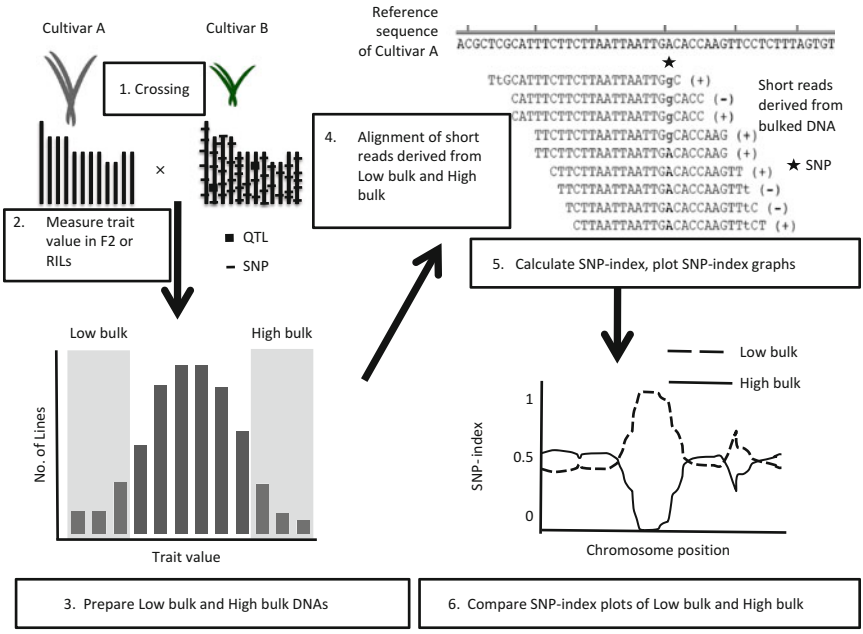

Fig. 3.2 A simplified scheme of QTL-seq. *1*. A cross is made between Cultivars A and B that show contrasting difference for the trait of interest. The chromosomes of the two cultivars are shown, and Cultivar A is used as reference. *2*. We score the phenotype of the progeny (F2 or RILs) derived from the cross between Cultivar A and B. *3*. If the trait is controlled by multiple QTL, the frequency distribution of trait value follows a Normal (Gaussian) distribution. We focus on the upper and lower tails of the distribution, and make two DNA-bulks, one from the upper tail (High bulk) and the other from the lower tail (Low bulk). *4*. High bulk and Low bulk DNAs are separately sequenced, and aligned to the reference sequence of either of the parent used for the cross (in this case Cultivar A). *5*. SNP-index is calculated and the relationships between SNP-index and chromosome position is depicted separately for High bulk and Low bulk. *6*. SNP-index graphs are compared between High and Low bulks. Genome positions with contrasting patterns of SNP-index plots between High and Low bulks indicate the localization of QTL differentiating the two bulks

in MutMap is calculated for all the SNPs identified between the each DNA-bulk and the reference, and the relationships between SNP-index and chromosome position (SNP-plots) are depicted separately for the High and Low bulks. For most of the genomic regions, genomes of the two parents are inherited to the progeny in equal probability; therefore, SNP-index should be around 0.5. However, genomic regions harboring QTL responsible for the differentiation of trait values among the progeny should exhibit contrasting patterns of SNP-index plots between High and Low bulks, which is easily identified by comparing SNP-index plots of the two bulks. We applied QTL-seq to F2 and RILs of rice, and successfully identified positions of QTL for partial resistance to blast fungus and seedling vigor. QTL-seq rapidly allows identification of QTL by two whole genome sequencing. Since all the SNPs in the genome are used as "genetic markers," the method is also applicable to progeny derived from crosses between closely related cultivars.

SNP-Index

In all the WGS-based methods described above, we used SNP-index for locating the candidate gene or genomic region. From the viewpoint of population genetics, SNP-index can be interpreted as a measure of nucleotide diversity of the population under study for a given genomic position. In the case of MutMap, the expected value of SNP-index in F2 population is 0.5, except for the genomic region harboring the causative mutation where SNP-index = 1. In bi-allelic situation, allele frequency of 0.5 corresponds to the highest genetic diversity, which is always reduced by deviating to 1 or 0. The SNP-index peak (SNP-index = 1) in MutMap can be viewed as a signature of selective sweep with reduced genetic diversity caused by selection of mutant type individuals in F2 population. Similarly, deviation of SNP-index values from 0.5 in QTL-seq is caused by selective sweep derived from phenotypic selection on the trait values (High and Low bulks). Therefore, we can reinterpret MutMap and QTL-seq as special applications of general methodology whereby SNP-index is used to identify genomic regions that underwent artificial selective sweeps. Note that MutMap and QTL are categorically "linkage studies" since we use progeny populations of F2 or RILs derived from known crosses. We expect that application of SNP-index-based method to "association study" will provide fruitful results not only in human but also in crop species in future.

Summary

Here we introduce a suite of WGS-based methods of gene/QTL identification in plants. We demonstrated that these methods can be applied to crop plants by applying them to rice with a focus on the improvement of an elite cultivar of Northern Japan. We expect that the MutMap and QTL-seq methods to have wide applicability in plant breeding along with related methods of mapping-by-sequencing that have been primarily developed in *Arabidopsis* (James et al. 2013).

The analysis pipelines are publicly available in the URL links below:
http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap

Acknowledgements This study was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences, the Ministry of Education, Cultures, Sports and Technology, Japan to HK and RT (Grant-in-Aid for Scientific Research on Innovative Areas 23113009) and JSPS KAKENHI to RT (Grant No. 24248004). We thank Shigeru Kuroda for general supports.

References

Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Yoshida K, Mitsuoka C, Tamiru M, Innan H, Cano L, Kamoun S, Terauchi R (2012) Genome sequencing reveals agronomically-important loci in rice from mutant populations. Nat Biotechnol 30:174–178

Avise JC (1994) Molecular markers, natural history and evolution. Chapman & Hall, New York
Bateson W, Saunders ER, Punnet MA (1905) Experimental studies in the physiology of heredity. Rep Evol Comm R Soc 2(1–55):80–99
Darvasi A, Soller M (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait loci. Genetics 138:1365–1373
Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics, 4th edn. Prentice Hall, London
Fekih R, Takagi H, Tamiru M, Abe A, Natsume S, Yaegashi H, Sharma S, Sharma S, Kanzaki H, Matsumura H, Saitoh H, Mitsuoka C, Utsushi H, Uemura A, Kanzaki E, Kosugi S, Yoshida K, Cano L, Kamoun S, Terauchi R (2013) MutMap+: genetic mapping and mutant identification without crossing in rice. PLoS One 8(7):e68529
Giovannoni JJ, Wing RA, Ganal MW, Tanksley SD (1991) Isolation of molecular markers from specific chromosome intervals using DNA pools from existing mapping populations. Nucleic Acids Res 19:6553–6558
International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800
James GV, Patel V, Nordström KJV, Klasen JR, Salomé PA, Weigel D, Schneeberger K (2013) User guide for mapping-by-sequencing in *Arabidopsis*. Genome Biol 14:R61
Lobo I, Shaw K (2008) Discovery and types of genetic linkage. Nat Educ 1(1):139
Mansur LM, Orf J, Lark KG (1993) Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (Glycine max L. Merr.). Theor Appl Genet 86:914–918
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. Proc Natl Acad Sci U S A 88:9828–9832
Morgan TH (1910) Sex-limited inheritance in *Drosophila*. Science 132:120–122
Rakshit S, Kanzaki H, Matsumura H, Rakshit A, Fujibe T et al (2010) Use of TILLING for reverse and forward genetics of rice. In: Meksem K, Kahl G (eds) The handbook of plant mutation screening: mining of natural and induced alleles. Wiley-VCH, Weinheim, pp 187–198
Shneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen J-E, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6:550–551
Takagi H, Uemura A, Yaegashi H, Tamiru M, Abe A, Mitsuoka C, Utsushi H, Natsume S, Kanzaki H, Matsumura H, Saitoh H, Yoshida K, Cano LM, Kamoun S, Terauchi R (2013a) MutMap-Gap: whole-genome resequencing of mutant F2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii. New Phytol 200(1):276–283. doi:10.1111/nph.12369
Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H, Cano L, Kamoun S, Terauchi R (2013b) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J 74:174–183

Chapter 4
Variant Calling Using NGS Data in European Aspen (*Populus tremula*)

Jing Wang, Douglas Scofield, Nathaniel R. Street, and Pär K. Ingvarsson

Introduction

The advent of next-generation DNA sequencing (NGS) technologies has revolutionized almost all biological disciplines by significantly increasing the speed and throughput capacities of data generation while simultaneously decreasing overall sequencing costs per gigabase (Metzker 2010). With its ability to tackle a set of challenges unconquered by traditional Sanger sequencing, NGS technology can remarkably advance both sequence-based genomic research and its downstream transcriptomic, epigenomic, or metagenomic studies, based on a broad range of sequencing methods, such as whole-genome, whole-exome, whole-transcriptome (RNA-seq), and chromatin immunoprecipitation (ChIP-seq) sequencing protocols; (Mardis 2008; Gilad et al. 2009; Frese et al. 2013). Most of these NGS-based studies almost inevitably depend on the accurate variants detection and genotype calling of variants from sequence data (Nielsen et al. 2011). Among the various types of variants detected through the whole-genome (re-)sequencing, single nucleotide polymorphisms (SNPs) and short insertions and deletions (indels) are the most abundant. However, currently, reliably calling SNPs and short indels and to infer genotype remain challenging despite that many recent computational algorithms

J. Wang, M.Sc. • P.K. Ingvarsson, M.Sc. (✉)
Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, Umeå 90187, Sweden
e-mail: par.ingvarsson@emg.umu.se

D. Scofield, Ph.D.
Department of Genetics and Evolution, Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

N.R. Street, Ph.D.
Department of Plant Physiology, Umeå Plant Science Centre, Umeå University, Umeå, Sweden

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_4

and statistical inference have been developed to both quantify and account for the large uncertainty associated with variant detection in NGS studies (Li et al. 2008, 2009b; McKenna et al. 2010).

In order to translate the raw sequencing data into final high-quality variant and genotype calls, a number of sophisticated computational analysis steps have to be carefully considered and performed. Meanwhile, in the past few years a flood of tools have been developed to handle the incredible amount of heterogeneous data produced during each step of the NGS data analysis workflow (Lee et al. 2012; Pabinger et al. 2013). All these computational processing, analyzing and interpreting of NGS data, and the appropriate choice of tools for each of the analysis steps present many obstacles and challenges for bench scientists. Here we firstly highlight some of the issues that may create ambiguities in variant and genotype calling, e.g. erroneous base calling, misalignment of sequence reads, skewed depths of read coverage. We also review and exemplify some recent methods and tools that are capable of improving the sensitivity and specificity of variant discovery from NGS data, including pre-processing of raw sequence reads, mapping reads to the reference genome, post-processing of the alignment results, variant discovery and genotype likelihood estimation, and finally the filtering out of false variants that represent artifacts of systematic sequencing errors or the result from inaccurate read alignments. We also provide guidelines for their application by using a data set of whole-genome re-sequencing data of 24 European aspen (*Populus tremula*) individuals each sequenced to a depth of about 20× coverage (Fig. 4.1) on an Illumina HiSeq 2000 sequencing platform.

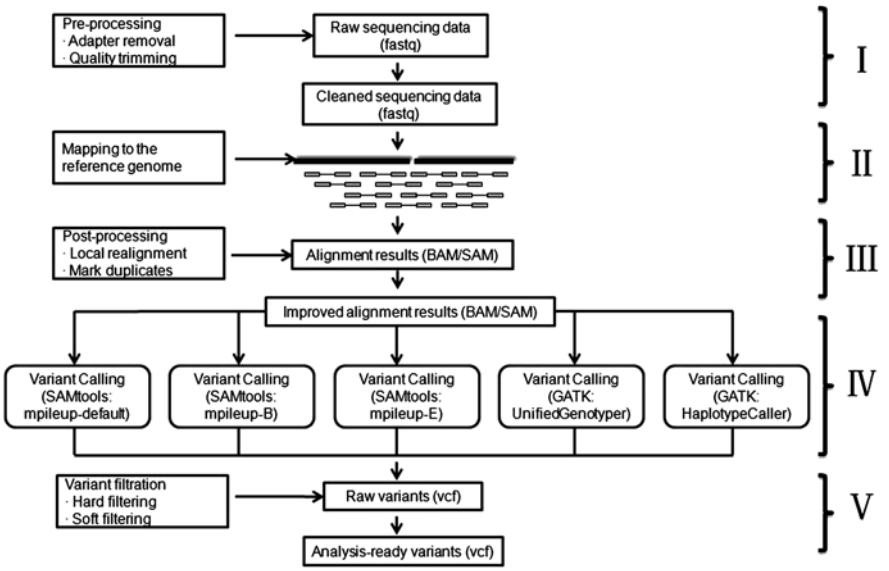

Fig. 4.1 Workflows for variant calling and genotype inference used in the whole-genome re-sequencing project of *P. tremula*

Raw Reads Pre-processing (Step I)

Compared to the Sanger sequencing technology, NGS data typically have a higher error rate which is caused by characteristics inherent of the NGS technologies, such as DNA damage, errors introduced during the process of amplification and sequencing (Liu et al. 2012). Such high sequencing error rates will significantly influence downstream read mapping, SNP and genotype calling since these algorithms all rely heavily on the sequencing error score (Bentley et al. 2008). Therefore, checking the quality of the raw sequence data is always the first step in any NGS analysis pipeline (Altmann et al. 2012; Minevich et al. 2012). It provides a quick overview on the base quality distributions and information on sequence properties such as possible base-calling errors, adapter contamination, or other sequence artifacts. Several tools have been available to produce general quality assessment reports, such as FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc), SolexaQA (Cox et al. 2010), and NGSQC Toolkit (Dai et al. 2010). Following quality assessment, the two most common procedures are adapter removal and quality trimming. Adapter contamination, the situation where adapter fragments are not completely removed from reads, will interfere with the downstream data analysis, and can lead to such problem as incorrect alignments of reads or even to erroneous variant calling. Hence, reads containing adapters must first be identified and then either trimmed or discarded prior to any following quality evaluation. Tools, such as Cutadapt (Martin 2011), AdapterRemoval (Lindgreen 2012), Btrim (Kong 2011), FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit), and Trimmomatic (Lohse et al. 2012), can all efficiently find and remove adapter sequences from NGS reads, although they have different features in identifying adapters in the 5′ or 3′ end of reads, searching for multiple different adapters simultaneously or processing single-end or paired-end sequencing data. Furthermore, it is well known that the quality of sequencing reads drops towards the end of reads, and therefore, it is always advisable to use Quality-based trimmers to trim off low-quality ends from reads that are likely to contain sequencing errors, in order to improve validity and accuracy in downstream analysis. Many quality control tools have been developed to perform flexible trimming and filtering tasks, almost all the aforementioned tools for trimming adapters can also be used to trim low-quality nucleotides from both ends of the reads where quality scores do not exceed a given threshold.

In our whole-genome re-sequencing project, we initially used Cutadapt to trim adapter sequences from reads generated on an Illumina HiSeq 2000 sequencing platform. On average, in 2.12 % of the processed reads an adapter occurred and these were subsequently trimmed. In additional, 0.14 % of the processed reads were completely discarded since their lengths were reduced below 36 bases after trimming (Fig. 4.2a). We then used Trimmomatic to cut bases off the start and end of each read when the quality dropped below 20, the program also scanned reads using a four-base pair wide sliding window and trimmed reads when the average quality per base dropped below 20. Finally it removed reads that were shorter than 36 base pairs after trimming. After the pre-processing of raw reads (step-I in Fig. 4.1), on average 83.3 % of paired reads survived the processing, and 4.6 % and 2.2 % of the reads only with either the forward or reverse read survived the processing, respectively (Fig. 4.2a).

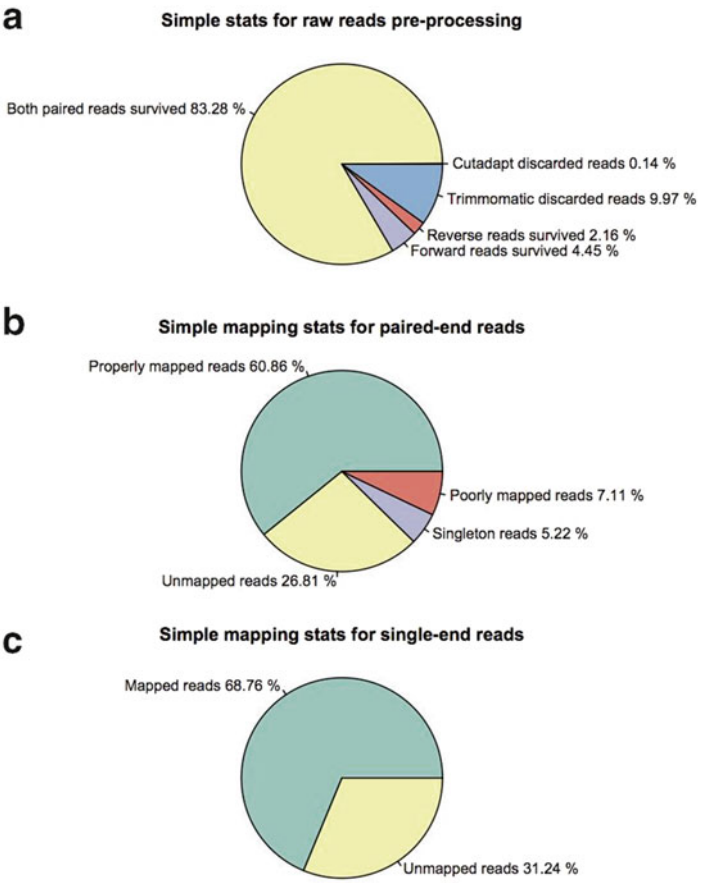

Fig. 4.2 Simple statistics of survived and trimmed reads after raw reads pre-processing (**a**) and mapping statistics that were reported by SAMtools flagstat for both paired-end (**b**) and single-end reads (**c**) in the whole-genome re-sequencing project of *P. tremula*

The Short-Read Alignment (Step II)

Following the pre-processing of raw reads, the second step in the accurate detection and inference of variants and genotypes is usually the mapping of sequence reads to a reference genome of either the sequenced or a closely related species (Flicek and Birney 2009). The accuracy of the alignment determines the ultimate outcome of variant calling, since wrongly aligned reads may produce artificial deviations from the reference, which will introduce errors in the downstream detection and analysis of genetic variants (Li et al. 2008). However, this task poses serious challenges due to huge amounts of sequence data, and it is usually both computationally intensive and time consuming (Bao et al. 2011). One challenge is that because reference

genomes are often very large, mapping millions or even billions of short reads against the genome within optimally memory and processing time usage is usually thought to be a computational bottleneck during the analysis of NGS data (Li and Homer 2010). Another commonly encountered obstacle is caused by repetitive DNA sequences, which are common throughout the genomes of a large number of species (Treangen and Salzberg 2011). Such repetitive sequences create ambiguities in alignment since some reads may map equally well to multiple locations in the genome. Furthermore, true variations between the sample and reference genomes and sequencing errors exacerbate this challenge, since they may even result in even more mismatches between the read and its true source compared to other copies of the repeat (Yu et al. 2012). In recent years, large numbers of algorithms and software tools have been developed to overcome these challenges, in order to accurately and efficiently map short reads to reference genomes. These tools include MAQ (Li et al. 2008), BWA (Li and Durbin 2009), SOAP2 (Li et al. 2009c), Bowtie (Langmead et al. 2009), NovoAlign (http://www.novocraft.com), and Stampy (Lunter and Goodson 2011). Here, we will emphasize some important features that need to be considered before initiating alignment and then briefly illustrate how to choose alignment tools aimed at specific features.

Alignment Algorithms

Most alignment algorithms use an indexing method to minimize processing time and they can be mainly grouped into two categories: Hash table-based (MAQ, NovoAlign) and Burrows Wheeler transform (BWT)-based method (BWA, Bowtie, SOAP2) (Li and Homer 2010). The advantages and limitations of different algorithmic approaches mainly reflect their efficiencies in tradeoff among mapping speed, memory usage, and mapping sensitivity and accuracy. It has been shown by recent studies that the BWT algorithm, which creates a complete reference genome index and then use parallelization strategies for alignment, can substantially reduced execution time and computer memory usage although at the cost of a slight loss of mapping sensitivity and accuracy compared to hash-based algorithms (Flicek and Birney 2009). Actually, the choice of an appropriate alignment method is strongly influenced by various factors, such as the specific research aims, the characteristics of the genome sequenced, the sequencing technology used, and the amount and length of the reads to be mapped.

Mismatches Between Sample and Reference Genome

Besides choosing an appropriate alignment algorithm, the parameter setting for each alignment tool is also essential for the accuracy of the alignment. The gap and mismatch penalty are important parameters one should carefully consider when

choosing an alignment tool (Bao et al. 2011). Allowing too few mismatches will obstruct downstream analysis in finding variants between sequenced and reference genomes. Conversely, allowing too much mismatches may promote more reads mapping to the reference genome but at the cost of increased number of false positive variants in the downstream analysis. Overall, the optimal choice of the gap opening and gap extension parameters and the tolerable number of mismatches highly depend on the species sequenced. Different alignment tools have different default options in setting these parameters, for example BWA can only align reads with at most a certain number of mismatches or gaps which depend on the length of the reads (Li and Durbin 2009), NovoAlign can allow up to eight mismatches per read for single end mapping (http://www.novocraft.com) whereas RMAP has no limitations on the number of mismatches (Smith et al. 2009).

Multiple Mapping

One of the most commonly encountered problems with the accuracy of the alignment are the large amounts of repetitive DNA sequences found in most species, which cause big problems for handling reads that map ambiguously to multiple locations in the reference genome and which therefore could lead to false inference of variants (Treangen and Salzberg 2011). Different alignment program treat repeats in different ways, one approach is to just ignore and discard all reads that can be mapped to multiple locations in the genome and to only report those reads that are uniquely mapped. Although suppressing reads with multiple alignments could greatly diminish the false alignments created by repetitive sequences, it might result in a substantial portion of biologically important variants being missed in the following analysis. Another commonly used strategy is the best match approach, in which only the alignment containing the fewest mismatches to the reference for a read is reported (e.g., BWA, Bowtie). While, if in cases where multiple equally best alignments exits, an aligner can either randomly choose one of them or discard all of them. Additionally, some alignment program will report all alignments for reads mapped to multiple positions (e.g., SOAP2). Furthermore, most state-of-the-art alignment tools (e.g., MAQ, BWA, Novoalign, Stampy) generate mapping quality scores for each alignment. This score is generally estimated by incorporating various factors, such as base qualities, the number of base mismatches, the existence and size of gap in the alignment, and it is often treated as an indicator of the likelihood that the alignment is accurate. The higher the mapping quality, the better the alignment is. A mapping quality of zero will be assigned to reads that can be aligned equally well to at least two positions. The accurate assessment of mapping quality for each read is very important since variant calling and other downstream analysis all depend on these scores (Ruffalo et al. 2011, 2012).

After aligning reads to a reference genome, most common alignment program store the alignment information in the Sequence Alignment/Map (SAM) format and/or its binary version, the Binary Alignment/Map (BAM) format (Li et al. 2009a).

These formats support almost all sequence types, store information about the position, orientation, and mapping quality of each aligned read and any paired read, and are efficient for access in the alignments in any specific regions. Moreover, many tools, e.g. SAMtools (Li et al. 2009a), Picard (http://picard.sourceforge.net), were developed to manipulate SAM/BAM alignment files, including sorting, merging, indexing, retrieving, and generating alignments in any genomic region swiftly. In addition, these tools can also compile statistics about the fraction of accurately mapped, poorly mapped, and unmapped reads for both single-end and paired-end reads. The visual inspection of alignments within a target genomic region can be performed using either a web-based genome browser or a stand-alone genome browser such as the Integrative Genomics Viewer (IGV) (Robinson et al. 2011).

In our whole genome re-sequencing project, we used the BWA aligner to align all paired-end and single-end reads to the reference genome and used SAMtools flagstat to generate a mapping statistics. On average, 60.9 % of all read pair were mapped the reference genome properly, whereas 7.1 % of total reads that have one read in a pair mapping on a different chromosome or with an unreasonable insert size. In 5.2 % of reads, one read in a pair was mapped but with the other read unmapped. Finally, 26.8 % of all reads could not be mapped to the reference at all (Fig. 4.2b). For single-end reads, 68.8 % of all reads mapped to the reference while the remaining were unmapped (Fig. 4.2c).

Post-processing Alignment (Step III)

Prior to the actual variant calling, multiple alignment post-processing steps, e.g. local realignment around indels, the marking of duplicated reads, and quality score recalibration need to be performed to minimize the number of artifacts in downstream variant calling.

Local Realignment Around Indels

In regions with insertions and deletions (indels), sequence reads are often mapped with mismatching bases effectively looking like evidence for variants. Sometimes the inconsistent placement of reads result in false positive or false negative variant calls, which further influence the accuracy of the downstream data analyses. Thus, local realignment can be performed to solve such alignment problems by first inspecting reads that overlap a given indel and then realigning individual reads to get a final consensus alignment. The local realignment tool in GATK (The Genome Analysis Toolkit) is one of the most common tools used to realign reads and yield more concise read mapping data in regions containing indels (McKenna et al. 2010; DePristo et al. 2011). It includes two steps, first the RealignerTargetCreator finds suspicious-looking intervals which are likely in need of realignment, and then the

IndelRealigner is used to run the realigner over those intervals. Except for the GATK tool, other tools such as Dindel (Albers et al. 2011) and SRMA (Short-Read Micro re-Aligner) (Homer and Nelson 2010) can also realign reads around small indels.

Mark Duplicates

Another commonly encountered issue is sequence duplication that is over-representation of certain sequences, which is an artifact of the PCR amplification step introduced during library construction. PCR duplicates often create a skewed coverage distribution that may subsequently bias the number of variants and substantially influence the accuracy of the variant discovery. If multiple reads or read pairs have identical external coordinates and the same insert length, several tools (e.g., MarkDuplicates in Picard and Rmdup in SAMtools) allow the user to either mark these duplicates or completely remove them, thereby only keeping the best quality reads out of all the identical reads. Compared to Rmdup in SAMtools that consider reads to be duplicates as long as their mapping locations are the same and does not work for unpaired reads (e.g., orphan reads or ends mapped to different chromosomes), MarkDuplicates in Picard seems to be a better choice for handling these cases since it not only takes into account mapping locations but also the sequence identity of reads aligned at those positions when marking or removing duplicates, and additionally it works properly for unpaired reads. Anyhow, removing duplicates not only can mitigate the effects of PCR amplification but also can reduce the potential source of noise and computational costs in downstream steps.

Base Quality Score Recalibration

When a sequencer calls a base, a quality score is always calculated by base-calling algorithms. However, many studies found that the raw Phred-scaled quality scores were often inaccurate and might deviate from the true base-calling error rates (Brockman et al. 2008). Thus, recalibrating base quality scores to make them more accurately reflect the true error rate is essential, as the variant and genotype calling in the downstream analyses all highly depend on the per-base quality scores (DePristo et al. 2011). One typical base recalibration procedure has been implemented in GATK, which first group bases into different categories with respect to several features, such as the reported quality score, the position of the base within the read, the dinucleotide context. Then, for each such category, the mismatch rate is computed and used to recalibrate the quality scores. However, it is often hard to run base quality score recalibration if there is not a comprehensive database of known SNPs, since the quality of the polymorphic sites will be inferred to be much lower than it actually is due to their mismatches to the reference sequence.

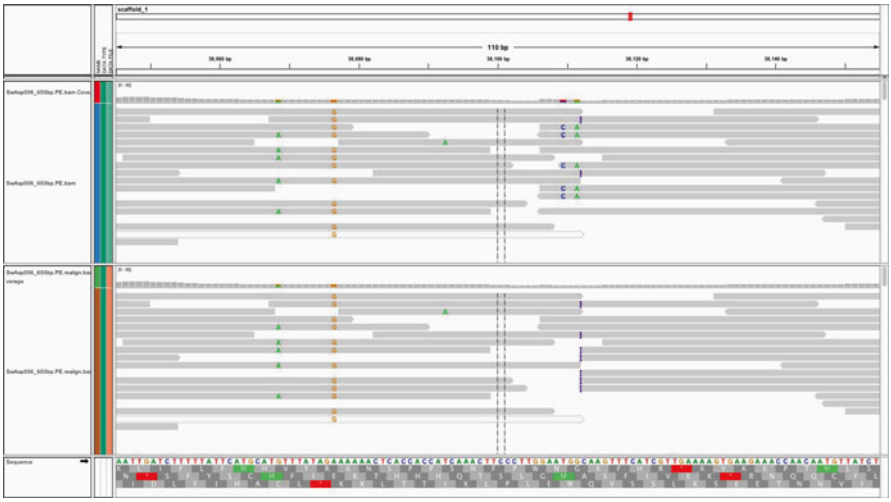

Fig. 4.3 Integrative genomics viewer (IGV) visualization of alignments in scaffold 1: 38,087–38,131 from the sample SwAsp006 before (*up*) and after (*down*) sequence realignment in the whole-genome re-sequencing project of *P. tremula*

In our whole genome re-sequencing project, we also carried out several processing steps to enhance the quality of the alignments before variant calling. First, initial alignments were refined by local realignment using GATK (Fig. 4.3), and then, we removed potential PCR duplicates using the Picard tool. From our data, it has been shown that there were relatively large amounts of reads which were marked as duplicates for both paired-end and single-end reads (Fig. 4.4). However, we did not recalibrate the base quality since we do not have a good SNP database for *P. tremula* yet.

Variant and Genotype Calling (Step IV)

One crucial, and nearly inevitable, application of NGS is variant detection and genotype calling at polymorphic sites. After aligning sequencing reads against a reference genome, variant calling can be performed by searching for positions where nucleotides from accession reads differ from the reference genome (DePristo et al. 2011; Nielsen et al. 2011). The relatively high error rates and short read length of NGS technologies present challenges for the accurate discovery of genetic variants and several empirical and statistical methods have therefore been developed to call variants (Li et al. 2008, 2009b; McKenna et al. 2010). Early variant calling approaches typically applied a serious of filtering steps according to thresholds on total read depth, per-base quality score, read alignment quality, strand-specific depth, and variant frequency (Wang et al. 2008). Genotype inference for each individual would then proceed by counting the number of supporting reads for each

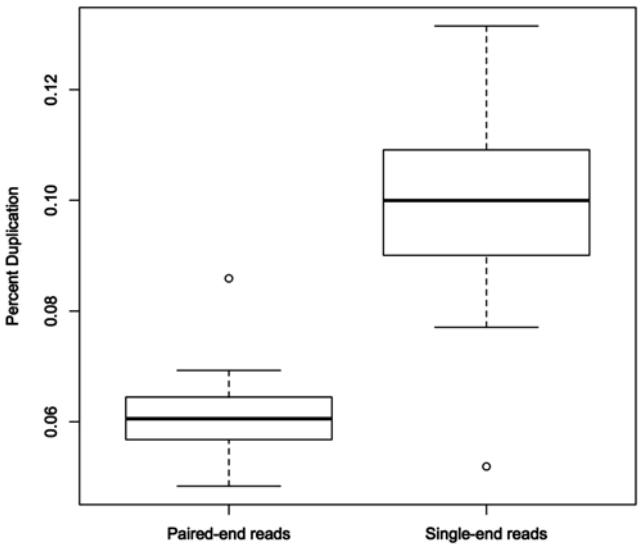

Fig. 4.4 The percentage of mapped sequence that is marked as duplicate for both paired-end reads and single-end reads in the whole-genome re-sequencing project of *P. tremula*

allele and using fixed cutoffs at the detected polymorphic sites. However, a main disadvantage of this type of genotype calling is that it would miss a lot of heterozygous genotypes in medium to low coverage regions. More recent variant calling methods use statistical models within a probabilistic framework and provide a natural way for measuring the uncertainty when calling variants and genotypes (Li et al. 2008, 2009b; McKenna et al. 2010; DePristo et al. 2011). In addition, the probabilistic methods can also incorporate information regarding allele frequencies across multiple individuals and patterns of linkage disequilibrium (LD) to improve the final genotype calling accuracy (Marchini and Howie 2010). In this section, we first focus on the methods used for generating genotype calls and calculating genotype likelihoods for all possible genotypes. We then provide an overview of some of the available software tools for variant discovery and genotyping.

Methods for Genotype Calling

Currently, the most common method for calling variants is estimating the probability of variant alleles through a Bayesian approach that integrates prior estimations of overall sequencing error rate and calculates the posterior probability of all the possible genotypes. With the genotype likelihood ($P(Data|G)$) which is computed from information such as the called nucleotides at each base, per-base quality values and mapping quality scores for each read, and prior probability for each genotype $P(G)$, the posterior probability of genotype G can be calculated via Bayes' formula.

The genotype with the highest posterior probability is often chosen as the most likely genotype call and this posterior probability can also be used as a measure of calling uncertainty.

Depending on how the prior probabilities for genotype are assigned, genotype calling methods can be classified into three categories: single-sample genotyping, multi-sample genotyping without integrating an LD-based analysis, and multi-sample genotyping using LD-based analysis (Nielsen et al. 2011; Li et al. 2012). For single-sample genotype calling method, the prior-genotype probability is either assigned equally for all genotypes or assigned differently based on external information such as different assignment strategies for homozygous and heterozygous genotypes, or depending on whether the variant has previously been reported in databases. Although the single-sample method performs well with high coverage NGS data, it produces large amounts of false positive variants with low coverage data (Li et al. 2009b). A number of multi-sample methods have therefore been developed to mitigate these issues by estimating allele frequencies from larger data sets, which can be further used to form more informative prior probabilities for inferring the true genotypes by integrating the Hardy–Weinberg equilibrium assumption from population genetic theory (Bansal et al. 2010; Martin et al. 2010). The genotype calling for the first two methods discussed so far are all performed at each site separately. However, integrating information of the pattern of LD at nearby sites can further substantially improve genotype calling accuracy when multiple samples have been sequenced (Scheet and Stephens 2006; Browning and Yu 2009). In practice, the choice of strategies to generate accurate genotype calls ultimately depends on the design of sequencing-based studies, especially on sequencing depth and the subsequent application of the data.

Software Tools for Variant Discovery and Genotyping

The upsurge of NGS methods has generated large amounts of data during the past few years and it is not surprising that a flood of methods and tools have been developed to contribute to accurate variant and genotype calling for NGS studies. Several widely used variant discovery tools, e.g. SAMtools (Li et al. 2009a), GATK (DePristo et al. 2011), and SOAP (Li et al. 2009c), all use a Bayesian statistical model that incorporates more realistic information on things such as read mapping quality, raw sequencing base quality, read depth, and the empirical correlation between adjacent qualities to simultaneously estimate the allele frequency and infer the genotype with highest probability at each site on both single sample and multi-sample data. In addition to Bayesian approaches, some tools such as SNVer use a more general frequentist framework for calling variants (Wei et al. 2011). SNVer does not require a prior probability as Bayesian methods since it formulates variant calling as a hypothesis-testing problem, and it can also call common and rare variants in both individual and pooled NGS data. Here, we give a brief overview of a number of commonly used tools for variant calling.

SAMtools mpileup and BCFtools

SAMtools is a versatile collection of tools for manipulating alignments in the SAM and BAM format. Variant and genotype calling by SAMtools are separated into two steps, SAMtools mpileup first computes the likelihood of all possible genotypes and stores these likelihood information in the BCF (Binary call format), which is the binary representation of the variant call format (VCF) (Danecek et al. 2011). Then, BCFtools from the SAMtools packages applies the prior and does the actual variant calling based on the genotype likelihoods calculated in the previous step. The advantage of the separation of genotype likelihood computation and subsequent variant calling in SAMtools enhances the flexibility for subsequent statistical analyses. Another important option in the SAMtools software package is the implementation of Base Alignment Quality (BAQ), which is a Phred-like score representing the probability that a read base is misaligned (Li 2011). The BAQ computation provides an efficient and effective way to reduce false SNP calls caused by misalignments around indels and dramatically improve the accuracy of SNP discovery. However, although the BAQ computation for each base is turned on by default in SAMtools mpileup, there have been recent suggestions that BAQ may be too strict and therefore leads to a lot of true SNPs being missed. It is feasible for users to disable probabilistic realignment for the computation of BAQ with the –B parameter, or perform an extended BAQ calculation with the –E option, which improves the sensitivity of variant discovery with relatively little extra cost of specificity.

GATK UnifiedGenotyper

The variant discovery pipeline of GATK may be the best state-of-the-art framework for variation discovery and genotyping using NGS data. The toolkit offers a wide variety of statistical models with a primary focus on discovering high-quality variant and genotype calls using multiple sequencing machines for many experimental designs. The GATK UnifiedGenotyper uses a similar Bayesian genotype likelihood model as SAMtools to simultaneously estimate the most likely genotypes and allele frequencies at each locus on both single and multi-sample data. In addition, the GATK provides an implementation of the BAQ algorithm from SAMtools, although BAQ is turned off by default in GATK. Another advantage of GATK is that it automatically applies several read filters before processing by UnifiedGenotyper or other pre- and post-processing steps, e.g. filtering out reads that fail the quality check, duplicate reads, mapping quality zero reads, reads whose mate maps to a different contig or unmapped reads.

SNVer

Differing from Bayesian methods, SNVer formulates variant discovery as a hypothesis testing problem, which reports one single overall p-value for evaluating the significance of a candidate locus being a variant in both pooled and individual NGS

data. Recently, a graphical user interface version of SNVer tool—SNVerGUI has been published (Wang et al. 2012). With SNVerGUI, the users can perform the entire variant calling pipeline by simply adjusting several parameters via a user-friendly graphical interface. Moreover, SNVerGUI supports all standard inputs and outputs and displays the results as tables, supporting various interactive post-call processing for further analysis.

GATK HaplotypeCaller

At present, the standard variant discovery and genotyping approach from NGS data is to map raw sequence reads to a reference genome and then identify positions where there exist simple variant sequences. This approach is well established and has been proven powerful for both single and multi-sample data. Nonetheless, the mapping-based approach does not consider the correlation between sequence reads when performing alignment and typically focuses only on a single variant type, which may result in inconsistent variant calling when different types of variants cluster. Several of these flaws can potentially be avoided through de novo assembly since it is agnostic with regard to variant type and divergence of sample sequence from the reference genome (Carnevali et al. 2012; Iqbal et al. 2012; Li 2012). Thus, assembly based variant calling is often treated as a complement to mapping based calling. The HaplotypeCaller from GATK call SNPs, indels and short structural variation (SV) simultaneously by performing a local de novo assembly of haplotypes via de Bruijn graphs in an active region if it has the potential to be variable, which is evaluated using an affine gap penalty Pair HMM model. If a variant call is determined and emitted, then a genotype for each sample will be assigned to it. The GATK HaplotypeCaller is still under active and continuous development and it currently only supports diploid calling.

In our whole-genome re-sequencing project, variant and genotype calling was performed with the two most widely used tools, SAMtools (version 0.1.18) and GATK (version 2.5-2). When calling variants using SAMtools mpileup, we tried three different options with regard to whether applying the BAQ algorithm: default option implements BAQ directly to evaluate the probability of misalignment of each base; option B disables the computation of BAQ; option E extends the BAQ computation. Additionally, we also used UnifiedGenotyper and HaplotypeCaller implemented in the GATK to call SNPs and short indels with near-default parameters except for the heterozygosity value which were used to compute prior likelihoods for both SNP and indel calling (The expected heterozygosity value was set to 0.015 for SNP calling and 0.0025 for indel calling).

A visual inspection of the variant calling using IGV showed that large amounts of SNPs that were called by all other methods, except for SAMtools with the default BAQ option, looked convincing in IGV (Fig. 4.5). Therefore, we did not use the default BAQ option in SAMtools in the following analysis. For the remaining four variant calling methods included in our study, we compared concordance and discrepancy of SNP calling between them when applied to the first 100 scaffolds for

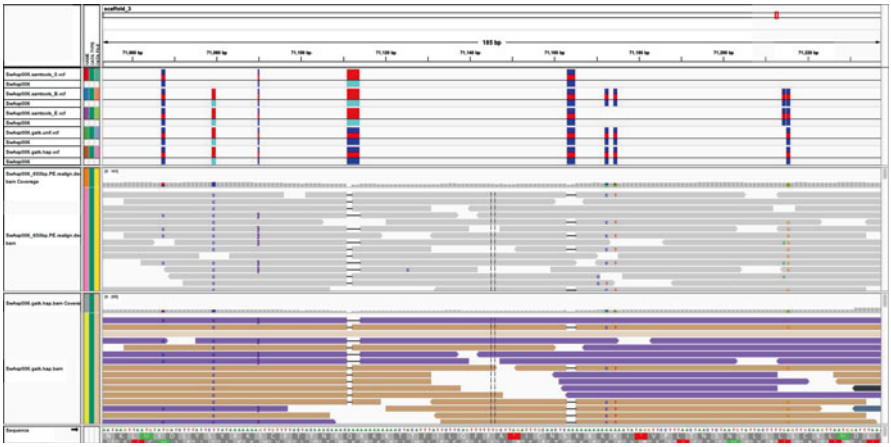

Fig. 4.5 Integrative genomics viewer (IGV) visualization of variants called by SAMtools default, SAMtools option B, SAMtools option E, GATK UnifiedGenotyper, GATK HaplotypeCaller (from *top* to *bottom*) in scaffold 3: 71,053–71,236 for the sample SwAsp006 in the whole-genome re-sequencing project of *P. tremula*. The HaplotypeCaller can also produce a BAM file to which assembled haplotypes were written and can be viewed through IGV (*bottom* in the IGV)

sample SwAsp006 (Fig. 4.6), and it showed a moderate discrepancy between the overall variant sets that might result from methodological variation between these different methods. Nonetheless, there still exist a set of robust SNPs that were called by all methods (representing nearly 70 % of all possible SNPs called) and these intersecting sites could therefore be regarded as true SNP data for the following filtering step. Although HaplotypeCaller is believed to be the best possible caller in GATK (http://www.broadinstitute.org/gatk/guide/topic?name=best-practices), we found a number of apparent errors with HaplotypeCaller with regard to how it chose candidate haplotypes and how it made heterozygous calls when performing variant calling in the version used (GATK version 2.5-2); again, this may be because we are working with a species with high heterozygosity. As a result, we chose UnifiedGenotyper to do SNP calling in our project, and the output of these raw SNP calls was later used in the filtering step.

Variant Filtration (Step V)

Most variant calling tools store DNA polymorphism data such as SNPs, indels, and structural variants in a standard VCF file (Danecek et al. 2011). The VCF file contains all variant information provided by the variant caller and usually includes eight fixed fields for each identified variant candidate: the chromosome (CHROM), a one-based position of the start of the variant (POS), unique identifiers of the variant (ID), the reference allele (REF), a comma separated list of alternate non-reference alleles

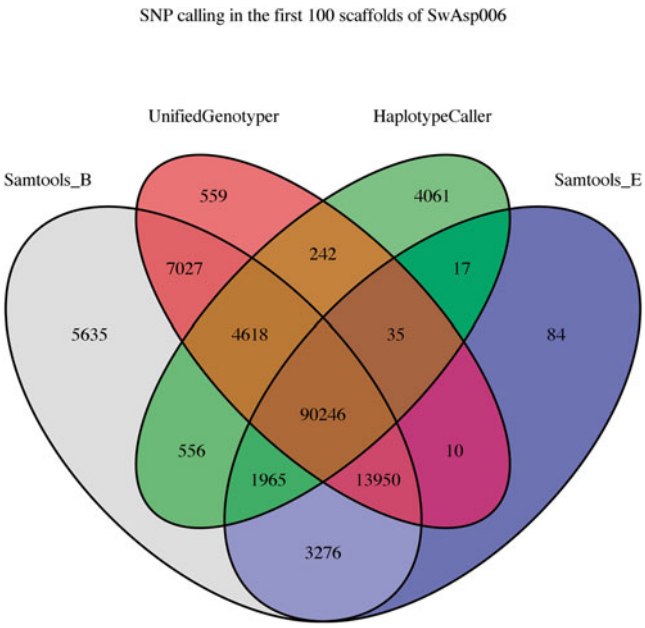

Fig. 4.6 Concordant and specific calls in SNP detection over the first 100 scaffolds between the four variant calling methods: SAMtools with option B, SAMtools with option E, GATK UnifiedGenotyper, GATK HaplotypeCaller for the sample SwAsp006 in the whole-genome re-sequencing project of *P. tremula*

(ALT), a phred-scaled quality score (QUAL), site filtering information (FILTER) and a semicolon separated list of additional, user extensible annotation (INFO). In addition, if samples and genotype information is present, the FORMAT column and sample IDs would follow the INFO field. Many tools have been developed to provide the possibility for easily processing VCF files. VCFtools is an open-source software package designed for working with VCF files, such as file validation, merging, intersecting, comparing, concatenating, filtering, and calculating some basic population genetic statistics, etc. (Danecek et al. 2011). Several tools in GATK also allow different operations to be performed on VCF files, e.g. CombineVariants and SelectVariants can merge or subtract VCF files from one another, VariantFiltration can filter variant calls using a number of user-selectable criteria, VariantRecalibrator and ApplyRecalibration can be combined to recalibrate variant quality score and then filter out false positive variants (DePristo et al. 2011). In addition, toolboxes such as BEDtools and SnpSift also include various utilities for manipulating VCF files (Quinlan and Hall 2010; Cingolani et al. 2012).

At the step of the variant calling process, the raw variant calls are likely to contain many false positives made from errors or incorrect alignment of the sequencing data, which can be greatly improved by using a number of filtering steps. Usually, the filtering approaches can be divided into two types: hard filtering and soft filtering (DePristo et al. 2011). Hard filtering is performed by

setting specific thresholds for variant properties and then removing all variants that do not fulfill these criteria. Typically, the most commonly used filters are those that remove genomic regions with repetitive DNA sequences where reads are poorly mapped to reference genome, or relies on filtering variant calls with properties outside normal distributions, such as extreme coverage depths, low confidence scores and mapping quality, strand bias, and so on. Hard filtering tools, such as VCFtools, GATK VariantFiltration and SnpSift can filter VCF files using arbitrary expressions as needed (Danecek et al. 2011; DePristo et al. 2011; Cingolani et al. 2012). However, the development and optimization of hard filters in terms of sensitivity and specificity are very difficult since they usually depend on various criteria, whereas the variants might meet the threshold in some criteria but not in others, and in this situation it is difficult to decide a variant should be filtered out or not. All of these challenges can be addressed with soft filtering, which do not need to apply any hand filters at any point in the process of variant filtration and all filtering criteria are learned from the data itself. In most cases, soft filtering is referred to as "variant quality score recalibration" (VQSR, performed in GATK) (DePristo et al. 2011). VQSR first creates a Gaussian mixture model given a set of putative high-quality variants along with covariant variant error annotations. This adaptive error model can then be employed to evaluate the probability that each variant call is real or not. Although VQSR was originally developed for variant calling in human re-sequencing data sets, it can also be used for other organisms since it is possible to create true variant data by experiments. Nonetheless, in some situations hard filtering has to be the only solution if we do not have an extensive variation database for our particular species.

In our whole-genome re-sequencing project, variant filtration was performed by combining the approaches of hard and soft filtering (Table 4.1). First, we called SNPs using GATK UnifiedGenotyper in the first 500 scaffolds with default options except for modifying the parameter of expected heterozygosity to 0.015 and 0.0025 for SNP and Indel calling, respectively. We chose to work with three independent samples which collected from the south, middle, and north of Sweden (SwAsp006, SwAsp078, SwAsp108). We found the total number of SNPs called were similar among individuals (Table 4.1). Then, we used hard filtering approach to remove genomic regions that contained at least two reads with mapping quality zero, thereby targeting variants in repetitive DNA regions. After this step, we found around 5 % SNPs were removed (Table 4.1). For the soft filtering approach, we first used the four variant discovery tools (SAMtools-B, SAMtools-E, GATK UnifiedGenotyper, GATK HaplotypeCaller)

Table 4.1 Number of SNPs initially called in the first 500 scaffolds for three samples (SwAsp006, SwAsp078, SwAsp108) and after initial hard filtering and subsequent soft filtering

	SwAsp006	SwAsp078	SwAsp108
Raw data	470,559	469,401	460,493
Hard filtering	443,839	443,193	434,022
Soft filtering	272,327	249,723	210,834

to call SNPs in genomic regions that contained in the 40–60 % quantiles of the total read coverage for each sample. We further extracted SNPs called by all these four tools using the GATK CombineVariants and SelectVariants and treated these SNPs as "true" SNPs and used these to develop a database to enable us to perform VQSR in GATK. Following VQSR, we found around 39 %, 44 %, 51 % SNPs were removed in SwAsp006, SwAsp078 and SwAsp108, respectively, which highly improve the accuracy and quality of SNP and genotype calling in our project.

Conclusion

Advances in NGS technology, coupled with the development of a large amount of novel, efficient statistical methods, and computational analysis tools, have made it possible to produce high-quality variant and genotype calls on an unprecedented scale. In this chapter we familiarized the reader with a number of analysis steps needed to perform towards optimizing the accuracy of variant and genotype calling. Additionally, we also presented a variant calling workflow starting from the raw sequenced reads to the filtration of the identified variants by processing a whole-genome re-sequencing dataset of a non-model species (*P. tremula*). Nonetheless, there still exist various challenges involved in obtaining accurate variant calls from NGS data. All of the currently available statistical algorithms, bioinformatic tools, and software for each of the optional steps in the variant calling workflow will rapidly evolve in response to changing sequencing platform technologies. As such, a recommended strategy for analyzing NGS data will change rapidly from month to month. Accordingly, rather than recommend any exact strategy, the chapter points out the various aspects need the user to consider when choosing suitable tools for optimal variant calling now or even in the future.

References

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R (2011) Dindel: accurate indel calls from short-read data. Genome Res 21(6):961–973

Altmann A, Weber P, Bader D, Preuß M, Binder EB, Müller-Myhsok B (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131(10):1541–1554

Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. Genome Res 20(4):537–545

Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y-Q (2011) Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet 56(6):406–414

Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18(5):763–770

Browning BL, Yu Z (2009) Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. Am J Hum Genet 85(6):847–861

Carnevali P, Baccash J, Halpern AL, Nazarenko I, Nilsen GB, Pant KP, Ebert JC, Brownley A, Morenzoni M, Karpinchyk V, Martin B, Ballinger DG, Drmanac R (2012) Computational techniques for human genome resequencing using mated gapped reads. J Comput Biol 19(3):279–292

Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X (2012) Using *Drosophila* melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Front Genet 3:35

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinform 11(1):485

Dai M, Thompson R, Maher C, Contreras-Galindo R, Kaplan M, Markovitz D, Omenn G, Meng F (2010) NGSQC: cross-platform quality analysis pipeline for deep sequencing data. BMC Genomics 11(Suppl 4):S7

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, Genomes Project Anal G (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498

Flicek P, Birney E (2009) Sense from sequence reads: methods for alignment and assembly. Nat Methods 6:S6–S12

Frese KS, Katus HA, Meder B (2013) Next-generation sequencing: from understanding biology to personalized medicine. Biology 2(1):378–398

Gilad Y, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. Trends Genet 25(10):463

Homer N, Nelson SF (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. Genome Biol 11(10):R99

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44(2):226–232

Kong Y (2011) Btrim: a fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. Genomics 98(2):152–153

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25

Lee HC, Lai K, Lorenc MT, Imelfort M, Duran C, Edwards D (2012) Bioinformatics tools and databases for analysis of next-generation sequence data. Brief Funct Genomics 11(1):12–24

Li H (2011) Improving SNP discovery by base alignment quality. Bioinformatics 27(8):1157–1158

Li H (2012) Exploring single-sample SNP and INDEL calling with whole-genome de novo assembly. Bioinformatics 28(14):1838–1844

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11(5):473–483

Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18(11):1851–1858

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009a) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J (2009b) SNP detection for massively parallel whole-genome resequencing. Genome Res 19(6):1124–1132

Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J (2009c) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25(15):1966–1967

Li Y, Chen W, Liu EY, Zhou Y-H (2012) Single nucleotide polymorphism (SNP) detection and genotype calling from massively parallel sequencing (MPS) data. Stat Biosci 1–23

Lindgreen S (2012) AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC Res Notes 5(1):337

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol. doi:10.1155/2012/251364

Lohse M, Bolger AM, Nagel A, Fernie AR, Lunn JE, Stitt M, Usadel B (2012) RobiNA: a user-friendly, integrated software solution for RNA-seq-based transcriptomics. Nucleic Acids Res 40(W1):W622–W627

Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21(6):936–939

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11(7):499–511

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24(3):133–141. doi:10.1016/j.tig.2007.12.007

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):10–12

Martin E, Kinnamon D, Schmidt M, Powell E, Zuchner S, Morris R (2010) SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. Bioinformatics 26(22):2803–2810

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303

Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11(1):31–46

Minevich G, Park DS, Blankenberg D, Poole RJ, Hobert O (2012) Cloudmap: a cloud-based pipeline for analysis of mutant genome sequences. Genetics 192(4):1249–1269

Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12(6):443–451

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, Trajanoski Z (2013) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. doi:10.1093/bib/bbs086

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011) Integrative genomics viewer. Nat Biotechnol 29(1):24–26

Ruffalo M, LaFramboise T, Koyutürk M (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics 27(20):2790–2796

Ruffalo M, Koyutürk M, Ray S, LaFramboise T (2012) Accurate estimation of short read mapping quality for next-generation genome sequencing. Bioinformatics 28(18):i349–i355

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629–644

Smith AD, Chung W-Y, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ (2009) Updates to the RMAP short-read mapping software. Bioinformatics 25(21):2841–2842

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13(1):36–46

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J (2008) The diploid genome sequence of an Asian individual. Nature 456(7218):60–65

Wang W, Hu WC, Hou F, Hu PZ, Wei Z (2012) SNVerGUI: a desktop tool for variant analysis of next-generation sequencing data. J Med Genet 49(12):753–755

Wei Z, Wang W, Hu PZ, Lyon GJ, Hakonarson H (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res 39(19):e132

Yu X, Guda K, Willis J, Veigl M, Wang Z, Markowitz S, Adams MD, Sun S (2012) How do alignment programs perform on sequencing data with varying qualities and from repetitive regions? BioData Min 5(1):1–12

Chapter 5
Leafy Spurge Genomics: A Model Perennial Weed to Investigate Development, Stress Responses, and Invasiveness

David Horvath, James V. Anderson, Wun S. Chao, Michael E. Foley, and Münevver Doğramaci

Introduction

Leafy spurge (*Euphorbia esula*, L.) is an invasive perennial weed (CABI 2004) that is related to several important crops including cassava (*Manihot esculenta*), castor bean (*Ricinus communis*), rubber tree (*Hevea brasiliensis*), and even more closely to poinsettia (*Euphorbia pulcherrima*), which is an important horticultural plant (Wurdack et al. 2005). Leafy spurge was introduced to North America in the late 1700s, probably in the ballast of ships. It was noted in several New England plant surveys shortly thereafter (CABI 2004). Although naturalized, it was never noted as a significant pest. However, in the late 1800s leafy spurge was introduced to the central USA, most likely by immigrants from Eastern Europe. It is speculated that Germans from Russia immigrants were the likely source for the introduction of leafy spurge to the Great Plains, and this hypothesis is supported by genetic analyses of leafy spurge populations in the USA and Eastern Europe (Rowe et al. 1995; Dunn 1985). It was speculated that the seeds were probably brought in as contaminants with grain (Dunn 1985), although this seems unlikely since leafy spurge is not generally found in cultivated plots. However, leafy spurge has been noted as a medicinal plant since ancient times—both as a regurgitant and as a skin exfoliant (Horvath et al. 2011). Thus, it seems possible that leafy spurge was brought here and cultivated as a folk remedy. It is these Eastern European strains that seem to dominate the gene pool of the most aggressive populations and have since spread throughout the Great Plains of the USA and Canada. By the late twentieth century, leafy spurge infested more than 2 million hectares and could be found in 35 US

D. Horvath, Ph.D. (✉) • J.V. Anderson, Ph.D. • W.S. Chao, Ph.D.
M.E. Foley, Ph.D. • M. Doğramaci, Ph.D.
Sunflower and Plant Biology Research Unit, USDA-ARS Red River Valley Agricultural Research Center, Bioscience Research Lab, 1605 Albrecht Blvd., Fargo, ND 58102, USA
e-mail: david.hovath@ars.usda.gov

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_5

states (Anderson et al. 2003; Quimby and Wendel 1997). Prior to the introduction of effective biological control agents, leafy spurge infested land was expanding at a rate between 12 and 16 % per year (Duncan et al. 2004), and leafy spurge was considered the most problematic weed in the Northern Great Plains (CABI 2004), costing land users over US$100 million annually (Leitch et al. 1994).

Leafy spurge infests primarily range, rights-of-ways, and recreation land in North America, and is a particular problem for some livestock producers since leafy spurge causes scours, severe irritation of the mouth and digestive system, and even death of the animals. Because leafy spurge is a rangeland weed, it is difficult to control by conventional means. However, goats and sheep will consume leafy spurge, and grazing practices combining cattle and sheep or goats have been devised as a means to control this noxious weed. Additionally, biological control with several introduced insects has proven effective in some locations. The most effective agent has been a group of flea beetles—specifically *Aphthona nigriscutis*, *Aphthona lacertosa*, *Aphthona czwalinae*, and *Aphthona flava* (Kirby et al. 2000). Although these cultural control measures and biological agents have shown some promise, there are still terrain features such as sandy soils, which can reduce the effectiveness of flea beetle control (Lym 2005), and remote areas that make distributing biological control agents difficult.

Leafy spurge is also fairly difficult to control using conventional herbicides. Several herbicide treatments have documented effectiveness (Lym and Messersmith 1983). However, because leafy spurge primarily infests areas that are of marginal agricultural value and often remote, it is often not cost effective to use chemical control measures. In addition, leafy spurge seeds can remain dormant in the ground for at least 8 years (Bowes and Thomas 1978), making chemical control a challenging task.

Underground adventitious shoot buds that form on the crown and roots (often referred to as crown and root buds) are another means that allows leafy spurge to avoid control measures. Leafy spurge has an extensive root system with hundreds of buds that will grow if the aerial portion of the plant is removed or destroyed disturbed (CABI 2004). These buds form seasonally on the crown and roots of leafy spurge, and once formed they remain in a paradormant state; that is, they will not grow unless they are separated from the growing meristems above them. Buds generally transition to a state of endodormant in the fall, which allows them to avoid initiation of new vegetative growth during brief warm spells in autumn after frosts have killed the above ground portion of the plant. Extended periods of cold temperature (i.e., vernalization) release buds from endodormancy and program predetermined floral meristems in preparation for growth-conducive temperatures (Anderson et al. 2005).

Because leafy spurge is such a difficult weed to control and results in considerable losses for US land users, in the early 1990s, the USDA Plant Science Research unit in Fargo, ND was redirected to study the biology of leafy spurge and to develop novel control measures. At the time, there was information available about the biology but limited or no information about the molecular biology and genetics of leafy spurge. The literature suggests that leafy spurge is an auto-allo hexaploid, usually

containing 60 chromosomes, although some variability in chromosome number was observed suggesting continued hybridization among leafy spurge subspecies (Stahevitch et al. 1988). Leafy spurge is self-fertile, but self-pollination produced fewer seeds than outcrosses—suggesting that there is some self-incompatibility (Selbo and Carmichael 1999).

Initial Forays into the Molecular Biology of Leafy Spurge

Studies using RFLPs designed from both nuclear and chloroplast DNA produced the first sequences cloned from leafy spurge (Rowe et al. 1995, 1997). However, the first sequenced genes of leafy spurge were derived from differential display experiments designed to identify genes induced when root buds of leafy spurge were cold treated (Horvath and Olson 1998), or to investigate release from endodormancy (Horvath and Anderson 2002). Several genes of known and unknown function were identified in these experiments. It was hypothesized that by examining the promoters of these genes, it might be possible to identify common *cis*-acting elements that could serve as binding sites for transcription factors required for the transition from a dormant to non-dormant state. Thus, the first genomic library for leafy spurge was created in a lambda zap express vector. This library had average insert sizes of about 6,000 bp and served as the source of genomic clones for many subsequent promoter cloning projects. Also, because differential display produces short gene fragments, a cDNA library was needed to clone and sequence full length transcripts. Thus, a cDNA library was created from mRNA of root buds that included both growth-induced and paradormant buds. In anticipation of future needs, this library was created in the hybrid-zap lambda vector in order to facilitate follow-on one and two-hybrid cloning experiments. Two reciprocally subtracted cDNA libraries from growing and dormant buds were also developed (Jia et al. 2006).

Research to clone genes involved in cell cycle, such as cyclins and *RETINO BLASTOMA* (*RB*) as well as genes involved in meristem growth and development such as *SHOOTMERISTEMLESS* (*STM*) was initiated. Primers designed to amplify *STM*, cyclins, histones, and cell cycle inhibitors such as the *KRP* family of genes proved reasonably effective (Anderson and Horvath 2001; Horvath et al. 2005a, b; Varanasi et al. 2008), and the cloning, sequencing, and characterization of full length clones from these genes from the cDNA and genomic libraries was usually effective. However, in some cases, the desired clones were not obtained.

Leafy Spurge Enters the Genomic Era

Cloning of several genes of interest had proven frustrating. Thus Dr. James Anderson employed the approach of randomly sequencing clones from cDNA library, which led to the establishment of the first EST-database for leafy spurge (Anderson and

Horvath 2001). Surprisingly, despite the cDNA library being neither normalized nor subtracted, the first 100 clones derived from this initial sequencing project included one of the histone genes that had proved difficult to clone using conventional methods. This initial EST-dataset contained ~1,100 unique cDNAs, which provided resources for monitoring the abundance of transcripts involved in response to xenobiotic-, cold-, and dehydration-stress (Anderson and Davis 2004), monitoring hybridization efficiency between leafy spurge and cassava, transcript abundance in response to heat-stress (Anderson et al. 2004), abundance of transcripts involved in carbohydrate metabolism, cell cycle, cell wall biochemistry, and response to auxin under changing environmental conditions (Anderson et al. 2005). As will be described later, this EST-dataset also provided the resources to develop the first leafy spurge microarrays, which were critical for allowing our group to explore and establish a functional genomics program for perennial weeds.

Our initial success led to development of a comprehensive EST-database, in collaboration with the International Institute for Tropical Agriculture, Ibadan, Nigeria, and the Keck Center at the University of Illinois, Urbana, USA. This project produced an EST-database for leafy spurge and an EST-database for cassava. The leafy spurge EST-database was developed from pooled RNAs from all plant tissues including dormant, after-ripened, and germinating seeds, as well as plant tissues exposed to cold, dehydration, wounding, photoperiod, and diseases such as downy mildew, and several insect biological control agents including gall midge and flea beetles. Because the leafy spurge library was normalized and subjected to three sequential subtraction procedures, the resulting 45,314 ESTs produced more than 23,000 unique gene sequences that collapsed into a 19,015 unigene set with an average size of 671 bases (Anderson et al. 2007). Two cassava EST-databases were also developed from RNA extracted from leaf tissues (cultivar TME117) of plants that were either well-hydrated or subjected to dehydration (Lokko et al. 2007). The cassava EST-database project generated 18,166 ESTs, which were used to identify 8,577 unigenes. Representatives of most cell cycle genes are present in these EST-databases as are genes involved in numerous hormone signaling responses and development programs. These EST-databases are available in Genbank, and the ESTIMA website (http://titan.biotec.uiuc.edu/cgi-bin/ESTWebsite/estima_start?seqSet=leafyspurge).

Experimenting with Transcriptomics

The promise and utility of cDNA microarrays for model plants such as arabidopsis (*Arabidopsis thaliana* Heyn) was well established by the turn of the century, but the use of these powerful techniques for examining gene expression in non-model plants was limited. We hypothesized that, since many of the most important genes are well conserved between plant species, it should be possible to follow the expression of many genes by hybridizing labeled cDNAs from non-model plants to arabidopsis-based cDNA microarrays. Thus, in collaboration with Michigan State

University, we were able to test this hypothesis. To our great surprise, we were able to visualize hybridization to nearly 80 % of the genes represented on the 7,000 element arabidopsis cDNA arrays using labeled leafy spurge cDNA (Horvath et al. 2003a). Even with cDNA from a weedy monocot (wild oat—*Avena fatua*) we were able to detect hybridization of more than 40 % of the genes on the arabidopsis arrays (Horvath et al. 2003b). More importantly, the hybridization was seemingly quite specific, as the test between cDNA derived from leaves verses those from growing meristems produced the expected differential expression of numerous cell cycle and photosynthesis genes expected to be differentially expressed between these tissues. However, subsequent cloning of the putative differentially expressed genes from leafy spurge often proved difficult since primers could only be designed from the available arabidopsis sequences for genes of interest. Also, the possibility of gene family members with differing expression patterns further complicated follow-on analyses.

With the assistance of Dr. Phil McClean at North Dakota State University, a small group of clones from our initial EST database were used to print the first leafy spurge microarrays. Although the arrays were useful tools in establishing procedures and served as a proof of concept, the complexity of the resulting arrays was recognizably insufficient for any significant analysis. Much better arrays were produced from a more complete set of ESTs by collaborators at the International Center for Tropical Agriculture, in Cali, Columbia. These arrays proved quite useful for initial investigations on paradormancy release (Horvath et al. 2005a, b) and other well-defined phases of dormancy in crown buds of leafy spurge (Horvath et al. 2006). However, even these arrays lacked the necessary complexity for a robust functional genomics analysis. Thus, once the large scale EST project was completed (Anderson et al. 2007), the University of Illinois produced microarrays from ~19,000 leafy spurge plus ~4,000 cassava unigenes that were not represented in the leafy spurge EST-database. Cassava unigenes were included because of our success with cross hybridizing arabidopsis cDNA arrays, and indications that many leafy spurge genes cassava can cross-hybridization (Anderson et al. 2004). These arrays proved to be one of the most important tools in our quest to develop a robust functional genomics program for understanding the biology of leafy spurge at a transcriptomic level.

Outcomes Obtained Using the 23,000 Element Leafy Spurge/Cassava Microarrays

The development and use of these microarrays produced an overwhelming abundance of results over several facets of leafy spurge biology and environmental responses. The arrays were first used to examine the differential gene expression during transitions in well-defined phases of seasonal dormancy in crown buds of leafy spurge under field conditions. This research identified nearly one thousand differentially expressed genes, and analysis of the data highlighted circadian

responses, hormones such as ethylene and abscisic acid, anoxia, and DORMANCY ASSOCIATED MADS-BOX transcription factors as playing a potential role in regulating different phases of dormancy induction and maintenance (Horvath et al. 2008). Further research established the impact of drought stress on endodormancy release (Doğramacı et al. 2011), the role of light and temperature on molecular mechanisms involved in well-defined phases of dormancy as well as flowering responses in underground adventitious buds of leafy spurge (Doğramacı et al. 2010, 2013), and the involvement of ethylene in the transition from para- to endodormancy (Doğramacı et al. 2013). These functional genomics projects have highlighted multiple gene products playing a role in dormancy processes in crown buds of leafy spurge, some of which are involved in circadian rhythm (CCA1, PIF3), photomorphogenesis (COP1, HY5), transcriptional regulation and stress response (AP2/ERF transcription factors), flowering (FT, MAF3) as well as many others associated with hormone signaling and response. Taken together, the development of EST-databases and incorporation of microarrays into our research program has allowed us to postulate and refine conceptual models for molecular mechanisms involved in regulating well-defined phases of seasonal dormancy in perennials (Horvath et al. 2002; Anderson et al. 2005; Doğramacı et al. 2010).

These arrays were also used to examine dormancy in seeds to follow changes in gene expression during transition from dormant- to germinating-seeds (Chao et al. 2011; Foley et al. 2010, 2011, 2013). These studies highlighted physiological responses in imbibed seeds that did not germinate at a constant temperature; leafy spurge seeds require day/night temperature fluctuations in order to efficiently germinate. These experiments again implicated circadian signals in dormancy responses with seeds similar to what was observed for buds and suggested a germination program that differs from of arabidopsis seeds.

With the assistance of Dr. Igor Andreev from the Ukrainian National Academy of Sciences, we also used these microarrays in a preliminary set of experiments designed to identify genes that were differentially expressed between leafy spurge plants growing in their native range near Kiev and to leafy spurge growing in its invaded range near Fargo, ND. Analysis of the data (Table 5.1) indicated that genes

Table 5.1 Selected results of significant ($p < 0.05$) processes identified through gene set enrichment and sub-network analysis from genes that were preferentially expressed in leafy spurge mature leaf tissue collected in the Ukraine vs. those that were preferentially expressed in US leafy spurge samples

Ukrainian spurge	U.S. spurge
Pathway up-regulated	
13-LOX and 13-HFL (jasmonic acid)	Starch biosynthesis
Systemin signaling	Carbohydrate biosynthetic process
Cell death	Starch biosynthetic process
Wounding response	Carbohydrate metabolic process
Salicylic acid stimulus response	Neighbors of RuBP carboxylase
Defense response	Neighbors of cyclin
Fungus response	Microtubule-based process

involved in biotic stress responses were more highly expressed in leafy spurge growing in the Ukraine relative to plants growing in the USA. These results suggest that leafy spurge has been released from diseases or insects that keep it in check in its native range, allowing it to grow more aggressively in its invaded range. Flea beetles were observed feeding on at least one of the four US samples and that another US sample was infested with a fungal pathogen. This observation suggests there are likely more devastating pathogens in the native range of leafy spurge, or that a specific agent is highly effective at inducing the defense response in its native range. It was also noteworthy that leafy spurge populations tested in the USA had greater abundance of transcripts associated with photosynthetic and carbon metabolism than did the leafy spurge growing in the Ukraine. This may be due to increased energy reserves being redirected from defense to production when leafy spurge is released from its pathogens. However, it could also be the results of genetic changes (or a combination of genetic and environmental difference) through evolution of leafy spurge in its invaded range. Common garden experiments are needed to test these hypotheses.

The response of leafy spurge to the bacteria that cause cassava blight (*Xanthomonas axonopodis* pv. *manihotis*) was also investigated in collaboration with Dr. Maria Santana from Universidad Simón Bolivar, Caracas, Venezuela. Interestingly, leafy spurge can be infected with this pathogen, with bacterial titers initially rising rapidly in the infected tissues (Horvath et al. 2013a, b). However, leafy spurge was usually able to shed infected leaves and recover, provided humidity levels were kept low. Correlations of early infection responses between leafy spurge and cassava were similar and indicated that there was an initial repression of photosynthesis-related genes. However, these genes rapidly resumed or exceeded normal expression levels in leafy spurge (Fig. 5.1), but most remained low or were further down-regulated in infected cassava plants—which eventually all died.

Because of our success with heterologous hybridization of arabidopsis microarrays, and the promising cross-hybridization between leafy spurge and cassava (Anderson et al. 2004) the leafy spurge arrays were also used to examine several physiological responses in cassava. These studies included tuberization and flowering between ancestral and domesticated varieties of cassava. Domestication of cassava likely involved evolutionary events related to removing wild ancestors from the rain forest and our studies indicate that the domestication process involves light signaling effects on chromatin remodeling, hormone balance, shoot branching, flowering, and tuber development involving asymmetric cell expansion resulting in reduced cell elongation and increased lateral root expansion (Anderson et al. 2012; Carvalho et al. 2009). The microarrays, as well as our EST-databases were even used to study protein production in normal and transgenic cassava (Carvalho et al. 2012; Horvath et al. unpublished), and to investigate changes in gene expression in poinsettia that were infected with a mycoplasm that alters auxin responses (Nicolaisen and Horvath 2008). Analysis still in progress from microarray studies also indicate that dehydration stress in related species of Euphorbiaceae (cassava, castor bean, leafy spurge) impacts common molecular

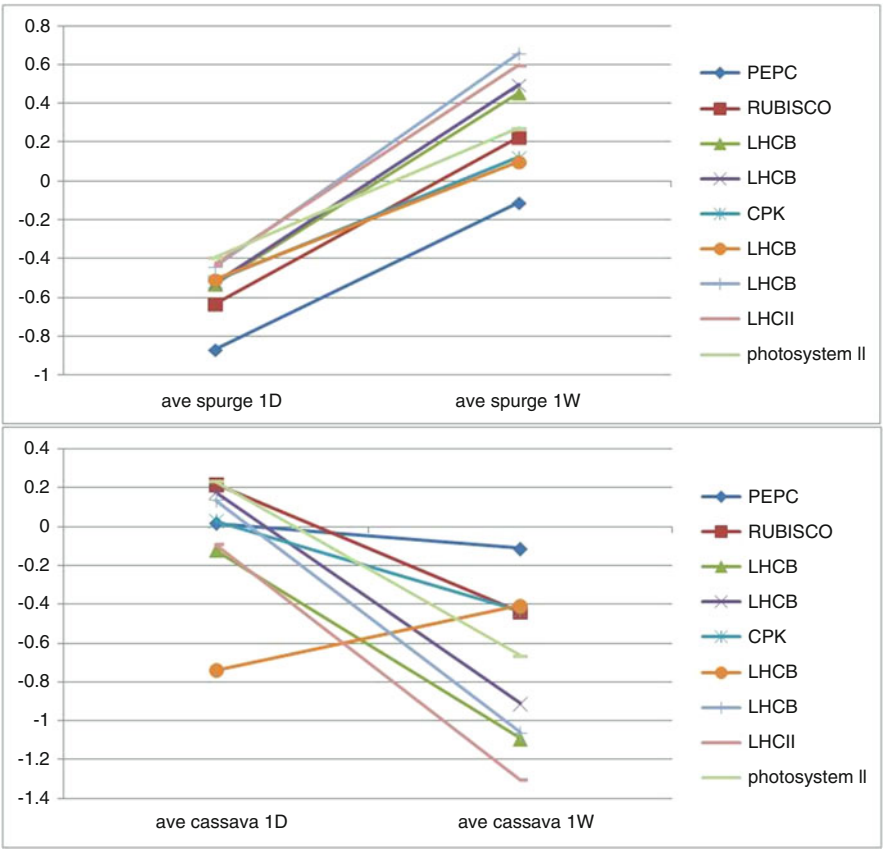

Fig. 5.1 Expression of genes associated with photosynthesis that were significantly down-regulated ($p<0.05$) in leafy spurge in response to pathogen within 1d. Graphs show expression of these genes in leafy spurge and cassava at 1d and 7d as the ratio between control and infected plants

mechanisms involved in circadian rhythm, jasmonic acid signaling and response, and photomorphogenesis (J.V. Anderson, unpublished), which is likely to provide new insights for maintaining crop productivity in response to the consequences of predicted global climate change.

BAC Library Construction: An Important Tool for Promoter Analysis and the First Step in Full Genome Sequencing

So far, well over 200 different microarrays have been hybridized and produced reliable data, which have been incorporated into bioinformatics programs to identify clusters of coordinately regulated genes from the various experiments. Thus, it is

possible to identify common sequences in the promoters of these coordinately regulated leafy spurge genes. Such sequences could serve as binding sites for transcription factors involved in various processes and responses to environmental conditions. To accomplish this, identification of the promoter sequences for these coordinately regulated genes is required. However, cloning of promoters from a genomics library is a laborious task that often required consecutive screenings to obtain clones with sufficient promoter sequences. Also, as leafy spurge is an auto-allo hexaploid, it was often impossible to know that contiguous sequences derived from several genomic clones represented a true gene sequence or if it was a chimera of several closely related paralogous genes. Additionally, on occasion, it was unclear if similar cDNAs were from two different genes or were the result of alternative splicing of the same gene. This was a conundrum in regard to the *DORMANCY ASSOCIATED MADS-BOX* (*DAM*) gene(s), which are known to be very similar to each other and also prone to alternative splicing. To deal with this issue, two separate bacterial artificial chromosome libraries (BAC) were created for leafy spurge—each with a different restriction enzyme. Each library contains 36,864 clones with an average size of 143 kb. The BAC libraries were developed in collaboration with the Arizona Genomics Institute and are publicly available. Combined, the libraries provided about 5× coverage of the 2.1 Gb leafy spurge genome (Horvath et al. 2013a, b). These libraries were used to clone multiple and highly similar *DAM* genes and confirmed that the two cDNAs resulted from alternative splicing of the same gene. We also identified and sequenced multiple members of the *FLOWERING LOCUS T* gene, which we hypothesize are potential targets for the DAM transcription factors.

Third generation sequencing technology from Pacific Biosciences (PacBIO) was initially used in an attempt to sequence selected BAC clones. Although substantial fold coverage was obtained, even when several BACs were sequenced simultaneously in the same reaction, the directional shifts and poor fidelity of the early PacBIO technology prevented easy assembly of the sequence. Consequently, these clones were sequenced by Amplicon Express, Pullman WA, from two different libraries with insert sizes of 200 and 400 bases using Illumina technology and their proprietary assembly pipeline. These processes generally resulted in 2–7 contigs per BAC. It would be interesting to utilize more recent PacBio assembly programs such as BLASR (BLASR website), perhaps after correcting the PacBio sequence data with the PacBioToCA program (Au et al. 2012), with the Illumina sequences generated by Amplicon Expressto determine if such efforts could produce single contig assemblies for each BAC. However, these procedures have not yet been attempted, since the major genes of interest were contained within single contigs produced by Amplicon Express.

In addition to serving as a source for large contiguous fragments of leafy spurge DNA, these BAC libraries could serve as the starting point to full genome sequencing of leafy spurge through a BAC by BAC cloning strategy. Plans are currently in place to develop a pooled matrix of both libraries, if funding becomes available. Using a 42×42×42 three dimension block matrix, all 73,728 clones can be represented in only 126 pools. These pools could be used to construct individually tagged Illumina libraries and sequenced on two lanes of Illumina, providing 15× coverage of the leafy spurge genome. An individual BAC clone would be represented by the intersection of three pools. Thus, it should be possible to de-convolute the sequence

for each BAC pool and devise a minimal tiling path for sequencing of the library. It may then be possible to fully assemble each BAC using a set of Illumina sequence data derived from shotgun sequencing (see below) of the whole leafy spurge genome using partially sequenced BAC contigs with programs such as IMAGE (Tsai et al. 2010) or PriceTI (Ruby et al. 2013) designed to close gaps. These sequenced and ordered BACs could finally result in a fully assembled genome for leafy spurge.

Shotgun Sequencing of the Leafy Spurge Genome

In 2012, we initiated several different approaches to yield a partial assembly of the leafy spurge genome using Illumina sequencing. We also investigated the possible use of third generation, long-read sequences generated by the PacBIO sequencing system. This is currently a work in progress.

The first sequences produced were from a 10 kb genomic library on four flow cells using the PacBIO C2 chemistry in collaboration with Dr. Tim Smith (USDA-ARS, Clay Center, NE USA). This research resulted in more than 150,000 reads, producing a total of 3.22^{e8} bases. The median contigs size (N50) was 3,073 bases and the maximum read length was 13,758 bases. Despite the fact that this represented only about 1/7th of the leafy spurge genome, approximately 1/3rd of our existing ESTs were represented in the PacBIO sequences and nearly 27,000 sequences had BlastN hits to the leafy spurge EST-database.

Subsequently, eight different genomic libraries were prepared from a group of genetically identical plants cloned from a single individual leafy spurge plant for Illumina sequencing. So far, four of these libraries have been sequenced; producing about 28× coverage of the leafy spurge genome and analyses of these sequences has been performed using iPlant resources (Goff et al. 2011).

Assembly of these sequences has been predictably problematic. Computer resources needed to run SOAP de novo on all four libraries in the iPlant discovery environment were insufficient. Thus, various techniques have been used to glean information from these sequences.

Because leafy spurge is closely related to two fully sequenced genomes (that of cassava and castor bean) and is in the same plant order (malpighiales) as poplar (*Populus* ssp.), the BLAT program was used to identify leafy spurge sequences that were conserved in one of these three fully sequenced species. The paired ends of each library were separately compared to each of the species, as well as to the leafy spurge EST database, and the resulting tabular data were combined in a single file. Duplicate entries were removed, and those sequences that had hits to at least one of the three fully sequenced genomes from at least one of the paired reads were used to build a "conserved fastq" file. Approximately 9 % of the leafy spurge fragments had hits to sequences from at least one of these three species. All fragments containing conserved sequences were combined and BLAT was again used to identify conserved repetitive sequences that were subsequently removed. The remaining conserved sequences represented approximately 8 % of the leafy spurge genome, which in theory contain the bulk of the gene sequences (Table 5.2).

Table 5.2 Read statistics from the four genomic libraries of leafy spurge

Library name	Average fragment size (bases)	Number of raw reads	Number of QC trimmed reads	Number of conserved reads	Number of repetitive elements on one or both sides	Percentage of conserved good reads	Percentage of good reads	Percentage of non-repetitive hits
Dorm 1	271	110,533,447	101,531,102	10,168,250	592,363	0.100149115	0.918555467	0.09431481
Dorm 2	403	72,286,512	65,715,399	6,992,799	231,409	0.106410356	0.909096278	0.10288897
Non Dorm 1	269	65,533,012	59,772,512	5,625,511	302,037	0.094115352	0.912097738	0.08906224
Non Dorm 2	373	80,687,457	73,999,074	7,767,140	274,953	0.104962665	0.917107525	0.10124704

Table 5.3 Statistics from the Trinity assembly of the conserved sequences

Count	Sum_len	N50	Min_len	Max_len	Med_len	Ave_len	SD_len
172,432	1.22E+08	886	201	19,081	594	707	484

These conserved sequences were assembled in the program Trinity and produced over 170,000 contigs with an average insert size of over 700 bases and cover about 1/20th of the leafy spurge genome (Table 5.3).

These contigs are expected to contain sequences primarily from recognized genes, but should lack promoter and intron sequences required for many analyses. However, a comparison of the resulting contigs to the previously sequenced BAC clones indicates that small introns and some 3′ and 5′ non-coding sequences were assembled.

Additional extension of these contigs will be attempted through several different approaches. One is to use the contigs generated by Trinity in a meta-assembly along with the combined Illumina libraries using an assembly program capable of handling long and short sequences such as MIRA (Chevreux et al. 2004) or Velvet (Zerbino and Birney 2008). Another possibility is to use iterative gap filling programs such as IMAGE or PriceTI to extend the ends of the resulting contigs. These programs extend seed sequences, such as the Trinity-produced contigs, by identifying and assembling fragments that map to the ends of the seed sequence and add them to the ends of the extending fragments. These programs then perform a meta-assembly to combine fragments that might now have overlapping sequences. The resulting assembly should represent the bulk of the conserved non-repetitive DNA in the leafy spurge genome. A comparison of selected fragments to previously sequenced BAC clones should confirm if this process provided an accurate assembly for most of the leafy spurge genes. Programs designed to identify and annotate genes can then be used to build a database of genes and associated promoter sequences.

Mining Old Data and New

A collection of contigs containing leafy spurge genes will provide the opportunity to extract additional data from the many microarray experiments previously done on leafy spurge. In most of the microarray analyses, clusters of genes coordinately regulated following particular treatment regimes were identified. Presumably these genes are coordinately regulated by common transcription factors. These transcription factors likely bind to similar *cis*-acting elements present in the regulatory regions of these coordinately regulated genes. Several programs are available that can identify over-represented short sequences within groups of promoter sequences. Such sequences are prime candidates for binding sites of transcription factors that regulate these genes in responses to specific conditions. Thus, it should be possible

to use these programs and the promoter sequences to identify transcription factors involved in seed and bud dormancy, environmental stress response, and even those that may have a functional role in the evolution of invasiveness.

A database of promoter sequences and the assembled gene space of leafy spurge will also be beneficial in future experiments using RNA sequencing (RNAseq). We have prepared libraries to RNAseq to re-examine the transcriptomics of paradormancy release and the response of leafy spurge to herbicide treatments. Additionally, antibodies to the DAM proteins of leafy spurge have been prepared and shown to specifically precipitate chromatin bound by these transcription factors. The availability of an assembled gene space makes possible the mapping of transcripts for accurate transcriptional analysis, and also for identifying genes that are regulated by DAM transcription factors with ChIPseq experiments.

In conclusion, years of effort have gone into making leafy spurge a model for understanding plant dormancy, with models for control of key dormancy-regulating gene networks and physiological processes. This research has required the building of genomic resources needed for observations and development of testable hypotheses regarding the mechanisms of leafy spurge invasiveness. Tools and techniques have been developed for assembling gene space and even whole genomes of complex non-model plant species at prices even moderately funded research programs can afford. This continuing research will make leafy spurge a true model system for complex, wild plant species.

References

Anderson JV, Davis DG (2004) Abiotic stress alters transcript profiles and activity of glutathione S-transferase, glutathione peroxidase, and glutathione reductase in *Euphorbia esula*. Physiol Plant 120:421–433

Anderson JV, Horvath DP (2001) Random sequencing of cDNAs and identification of mRNAs. Weed Sci 49:590–597

Anderson GL, Delfosse ES, Spencer NR, Prosser CW, Richard RD (2003) Lessons in developing successful invasive weed control programs. J Range Manage 56:2–12

Anderson JV, Delseny M, Fregene MA, Jorge V, Mba C, Lopez C, Restrepo S, Soto M, Piegu B, Verdier V, Cooke R, Tohme J, Horvath DP (2004) An EST resource for cassava and other species of Euphorbiaceae. Plant Mol Biol 56:527–539

Anderson JV, Gesch RW, Jia Y, Chao WS, Horvath DP (2005) Seasonal shifts in dormancy status, carbohydrate metabolism, and related gene expression in crown buds of leafy spurge. Plant Cell Environ 28:1567–1578

Anderson JV, Horvath DP, Chao WS, Foley ME, Hernandez AG, Thimmapuram J, Liu L, Gong GL, Band M, Kim R, Mikel MA (2007) Characterization of an EST database for the perennial weed leafy spurge: an important resource for weed biology research. Weed Sci 55:193–203

Anderson JV, Carvalho LJCB, de Souza CRB, Vieira EA (2012) Genomic studies of genetic diversity between *Manihot esculenta* spp. *esculenta* and its ancestor *Manihot esculenta* spp. *flabellifolia* reveals regulatory pathways related to storage root formation and flowering behavior. In: Second scientific conference of the Global Cassava Partnership for the 21st century, June 18–22, Speke Resort and Conference Centre, Kampala, Uganda

Au KF, Underwood JG, Lee L, Wong WH (2012) Improving PacBio long read accuracy by short read alignment. PLoS One 7:e46679

BLASR website. Available: http://www.smrtcommunity.com/SMRT-Analysis/Algorithms/BLASR. Accessed 8 Sept 2012

Bowes GG, Thomas AG (1978) Longevity of leafy spurge seeds in the soil following various control programs. J Range Manage 31:137–140

CABI (2004) *Euphorbia esula* (original text by Chao W and Anderson JV). In: Crop Protection Compendium, 2004 edition. CAB International, Wallingford

Carvalho LJCB, de Souza CRB, de Mattos Cascardo JC, Valle Agostini MA, Alano Vieira E, Anderson JV, Lippolis J (2009) Natural genetic variation in cassava (*Manihot esculenta* Crantz) landraces: a tool for gene discovery. In: Shu QY (ed) Induced plant mutations in the genomics era. Food and Agriculture Organization of the United Nations, Rome, pp 313–316. ISBN 978-92-5-106324-8

Carvalho LJCB, Lippolis J, Chen S, de Souza CRB, Vieria EA, Anderson JV (2012) Characterization of carotenoid-protein complexes and gene expression analysis associated with carotenoid sequestration in pigmented cassava (*Manihot esculenta* Crantz) storage root. Open Biochem J 6:116–130

Chao WS, Foley ME, Doğramacı M, Anderson JV, Horvath DP (2011) Alternating temperature breaks dormancy in leafy spurge seeds and impacts signaling networks associated with HY5. Funct Integr Genomics 11:637–649

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE 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

Doğramacı M, Horvath DP, Chao WS, Foley ME, Christoffers MJ, Anderson JV (2010) Extended low temperature impacts dormancy status, flowering competence, and transcript profiles in crown buds of leafy spurge. Plant Mol Biol 73:207–226

Doğramacı M, Horvath DP, Christoffers MJ, Anderson JV (2011) Dehydration and vernalization treatments identify overlapping molecular networks impacting endodormancy maintenance in leafy spurge crown buds. Funct Integr Genomics 11:611–626

Doğramacı M, Foley ME, Chao WS, Christoffers MJ, Anderson JV (2013) Induction of endodormancy in crown buds of leafy spurge (*Euphorbia esula* L.) implicates a role for ethylene and cross-talk between photoperiod and temperature. Plant Mol Biol 81:577–593

Duncan CL, Jachetta JJ, Brown ML, Carrithers VF, Clark JK, DiTomaso JM, Lym RG, McDaniel KC, Renz MJ, Rice PM (2004) Assessing the economic, environmental, and societal losses from invasive plants on rangeland and wildlands. Weed Technol 18:1411–1416

Dunn PH (1985) Origins of leafy spurge in North America. In: Watson AK (ed) Leafy spurge, monograph no. 3. Weed Science Society of America, Champaign, pp 7–13

Foley ME, Anderson JV, Chao WS, Doğramacı M, Horvath DP (2010) Initial changes in the transcriptome of *Euphorbia esula* seeds induced to germinate with a combination of constant and diurnal alternating temperatures. Plant Mol Biol 73:131–142

Foley ME, Chao WS, Doğramacı M, Horvath DP, Anderson JV (2011) Changes in the transcriptome of dry leafy spurge seeds imbibed at a constant and alternating temperature. Weed Sci 60:48–56

Foley ME, Chao WS, Horvath DP, Doğramacı M, Anderson JV (2013) The transcriptomes of dormant leafy spurge seeds under alternating temperature are differentially affected by a germination-enhancing pretreatment. J Plant Physiol 170:539–547

Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N et al (2011) The iPlant collaborative: cyberinfrastructure for plant biology. Front Plant Sci 2:34

Horvath DP, Anderson JV (2002) A molecular approach to understanding root bud dormancy in leafy spurge. Weed Sci 50:227–231

Horvath DP, Olson PA (1998) Cloning and characterization of cold-regulated glycine-rich RNA-binding protein genes from leafy spurge (*Euphorbia esula* L.) and comparison to heterologous genomic clones. Plant Mol Biol 38:531–538

Horvath DP, Schaffer R, West M, Wisman E (2003a) Arabidopsis microarrays identify conserved and differentially expressed genes involved in shoot growth and development from distantly related plant species. Plant J 34:125–134

Horvath DP, Schaffer R, Wisman E (2003b) Identification of genes induced in emerging tillers of wild oat (*Avena fatua*) using Arabidopsis microarrays. Weed Sci 51:503–508

Horvath DP, Anderson JV, Jia Y, Chao WS (2005a) Cloning, characterization, and expression of growth regulator CYCLIN D3-2 in leafy spurge (*Euphorbia esula*). Weed Sci 53:431–437

Horvath DP, Soto-Suárez M, Chao WS, Jia Y, Anderson JV (2005b) Transcriptome analysis of paradormancy release in root buds of leafy spurge (*Euphorbia esula*). Weed Sci 53:795–801

Horvath DP, Anderson JV, Soto-Suarez M, Chao WS (2006) Transcriptome analysis of leafy spurge (*Euphorbia esula*) crown buds during shifts in well-defined phases of dormancy. Weed Sci 54:821–827

Horvath DP, Chao WS, Suttle JC, Thimmapuram J, Anderson JV (2008) Transcriptome analysis identifies novel responses and potential regulatory genes involved in seasonal dormancy transitions of leafy spurge (*Euphorbia esula* L.). BMC Genomics 9:536

Horvath D, Wurkdack K, Pullin KL (2011) Euphorbia. In: Kole C (ed) Wild crop relatives: genomic and breeding resources. Springer, Berlin, pp 125–132

Horvath DP, Kudrna D, Talag J, Anderson JV, Chao WS, Wing RA, Foley ME, Doğramacı M (2013a) BAC library development, and clone characterization for dormancy-responsive DREB4A, DAM, and FT from leafy spurge (*Euphorbia esula* L.) identifies differential splicing and conserved promoter motifs. Weed Sci 61:303–309

Horvath DP, Santana M, Anderson JV (2013b) Microarray analysis of the semicompatible, pathogenic response and recovery of leafy spurge (*Euphorbia esula*) inoculated with the cassava bacterial blight pathogen *Xanthomonas axonopodis* pv. *manihotis*. Weed Sci 61:428–436

Jia Y, Anderson JV, Horvath DP, Gu Y-Q, Lym RG, Chao WS (2006) Subtractive cDNA libraries identify differentially expressed genes in dormant and growing buds of leafy spurge (*Euphorbia esula*). Plant Mol Biol 61:329–344

Kirby DR, Carlson RB, Krabbenhoft KD, Mundal D, Kirby MM (2000) Biological control of leafy spurge with introduced flea beetles (*Apththona* spp.). J Range Manage 53:305–308

Leitch JA, Leistritz FL, Bangsund DA (1994) Economic effect of leafy spurge in the Upper Great Plains: methods, models, and results. North Dakota State University Agric Econ Rep 316, 8 pp

Lokko Y, Anderson JV, Rudd S, Raji A, Horvath D, Mikel MA, Kim R, Liu L, Hernandez A, Dixon AGO, Ingelbrecht I (2007) Characterization of an 18,166 EST dataset for cassava (*Manihot esculenta* Crantz) enriched for drought-responsive genes. Plant Cell Rep 26:1605–1618

Lym RG (2005) Integration of biological control agents with other weed management technologies: successes from the leafy spurge (*Euphorbia esula*) IPM program. Biol Control 35:366–375

Lym RG, Messersmith CG (1983) Control of leafy spurge with herbicides. N D Farm Res Bull 40:16–19

Nicolaisen M, Horvath D (2008) A branch-inducing phytoplasma in *Euphorbia pulcherrima* associated with changes in expression of host genes. J Phytopathol 156:403–407

Quimby PC Jr, Wendel L (1997) The ecological area-wide management (TEAM) – leafy spurge. Sidney, MT. In Executive Summary, USDA, ARS, Wide Area funding proposal, p 51

Rowe M, Lee D, Nissen S, Masters R, Lee D, Bowditch M (1995) Relatedness of North American and European leafy spurge based on DNA markers. Leafy spurge symposium. Fargo, ND, July 25–27, 1995, p 33

Rowe M, Lee D, Nissen S, Bowditch B, Masters R (1997) Genetic variation in North American leafy spurge (*Euphorbia esula*) determined by DNA markers. Weed Sci 45:446–454

Ruby JG, Bellare P, Derisi JL (2013) PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. G3 – Genes Genomes Genet 20:865–880

Selbo SM, Carmichael JS (1999) Reproductive biology of leafy spurge (*Euphorbia esula* L.): breeding system analysis. Can J Bot 77:1684–1688

Stahevitch AE, Crompton CW, Wojtas WA (1988) Cytogenetic studies of leafy spurge, *Euphorbia esula* and its allies (Euphorbiaceae). Can J Bot 66:2247–2257

Tsai IJ, Otto TD, Berriman M (2010) Method improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol 11:R41

Varanasi V, Slotta T, Horvath D (2008) Cloning and characterization of a critical meristem developmental gene (EeSTM) from leafy spurge (*Euphorbia esula*). Weed Sci 56:490–495

Wurdack KJ, Hoffmann P, Chase MW (2005) Molecular phylogenetic analysis of uniovulate Euphorbiaceae (*Euphorbiaceae sensu stricto*) using plastid RBCL and TRNL-F DNA sequences. Am J Bot 92:1397–1420

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Chapter 6
Utilization of NGS and Proteomic-Based Approaches to Gain Insights on Cellular Responses to Singlet Oxygen and Improve Energy Yields for Bacterial Stress Adaptation

Roger S. Greenwell Jr., Mobashar Hussain Urf Turabe Fazil, and H.P. Pandey

Introduction

Cellular responses to environmental cues dictate the ability of organisms to adapt and survive. Organisms will reprogram their metabolic activities in response to these cues, altering a wide number of biological traits such as to increase the likelihood of passing on their genetic material. One of the simplest primary responses are transcriptional response mechanisms, as a wide variety of target genes can be differentially expressed based on necessity, and can be regulated by a variety of transcription factors. The net effect of these signal transduction pathways involves changes in growth, metabolism, motility, and transcriptional and translational activity. For example, microbes have evolved responses to molecular O_2 that dictate cellular growth between aerobic and anaerobic states while also contending with any toxic by-products generated as a consequence of those growth conditions.

Indeed, as the first photosynthetic organisms acquired the ability to produce significant quantities of O_2, atmospheric accumulation led to the evolution of bioenergetic pathways. The evolution of aerobic respiration, wherein the formation of a

R.S. Greenwell Jr., Ph.D. (✉)
Biology Department, Worcester State University,
486 Chandler Street, Worcester, MA 01602, USA
e-mail: rgreenwell@worcester.edu

M.H.U.T. Fazil, B.Sc., M.Sc., Ph.D. (✉)
Dermatology and Skin Biology, Lee Kong Chian School of Medicine,
50 Nanyang Drive, Research Techno Plaza, Level 4, X-Frontiers Block,
Singapore 637553, Singapore
e-mail: fazil.turabe@gmail.com; turabe.fazil@ntu.edu.sg

H.P. Pandey, Ph.D.
Faculty of Science, Department of Biochemistry, Banaras Hindu University,
Varanasi, UP, India

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_6

proton gradient to generate ATP is coupled with reduction of O_2, led to increased biological diversity through eventual adaptation and development of eukaryotic organisms (Ziegelhoffer and Donohue 2009; Kerr 2005). The use of O_2 as a terminal electron acceptor conferred a growth advantage in part due to the large amount of conserved energy in the reduction to water, done via a 4-electron transfer reaction by cytochrome oxidases (Miller and Gennis 1986; Azzi and Gennis 1986). However, one important trade-off to using O_2 as an electron acceptor is the formation of various reactive oxygen species (ROS) that can cause severe damage and lead to cell death (Rosner and Storz 1997; Schulz et al. 2000c).

While molecular O_2 is relatively inert due to the spin state restriction, it is converted to toxic ROS by either one electron transfer (class I) or energy transfer (class II) reactions (Fig. 6.1) (Steinberg 2012; Kiley and Storz 2004; Ziegelhoffer and Donohue 2009). These compounds can significantly damage a wide variety of biomolecules, triggering the onset of debilitating diseases or leading to cell death (Cogdell et al. 2000; Schulz et al. 2000c; Frank and Brudvig 2004b). Significant effort has been invested in determining the effects of and responses to the presence of the class I ROS, but less information has been gathered regarding response to the

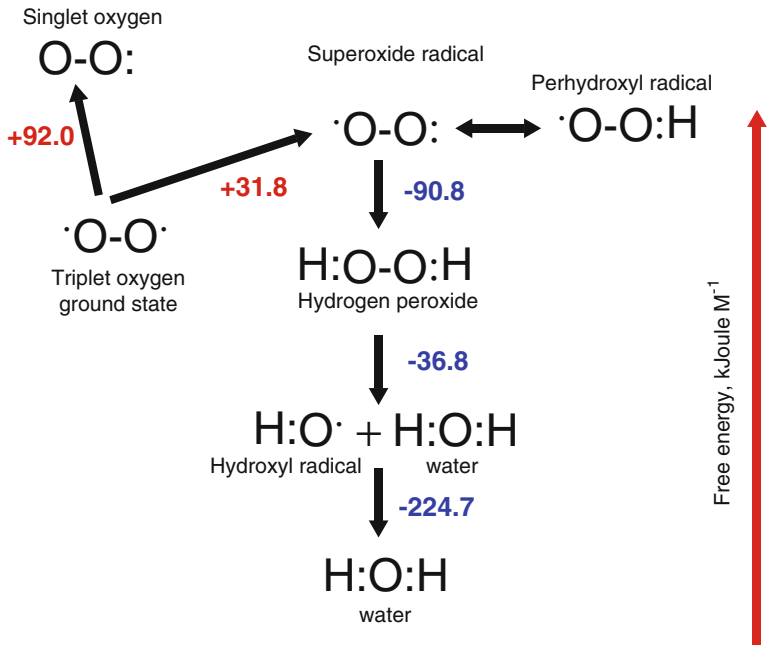

Fig. 6.1 Generation of reactive oxygen species. Various forms and energy of generation for reactive oxygen species. *Left*: Energy (92 kJ M^{-1}) required to activate the triplet state into class II ROS, singlet oxygen. *Right*: After the endergonic (31.8 kJ M^{-1}) reduction of O_2 to superoxide, the subsequent reduction steps (to hydrogen peroxide, hydroxyl radicals, and lastly complete reduction to water) are exergonic and occur spontaneously. *Red* denotes endergonic reactions, *blue* denotes exergonic reactions (Steinberg 2012)

class II ROS (Schulz et al. 2000a, b, c; Kiley and Storz 2004; Storz and Imlay 1999; Zheng and Storz 2000).

This chapter will discuss the generation of and cellular responses to the class II ROS, singlet oxygen ($^1O_2$). Particular focus will be on the effects of $^1O_2$ on the purple non-sulfur photoheterotroph, *Rhodobacter sphaeroides*, and the cellular response triggered by *R. sphaeroides* in the presence of this ROS. In-depth reviews on the responses to singlet oxygen in microbes can be found elsewhere (Ziegelhoffer and Donohue 2009; Glaeser et al. 2011). We will discuss the generation of and damage caused by $^1O_2$, and the triggering of the transcriptional response cascade in response to this ROS. We will focus on the use of next-generation sequencing technologies and multi-omics experiments for insights on this response and the utilization of such technologies toward improved energy generation.

Reactive Oxygen Species and Singlet Oxygen

The relatively stable ground state of oxygen is a triplet state with two unpaired electrons with the same spin quantum number, each located in different antibonding (π*) orbitals. Oxygen can react by oxidizing another molecule, but, despite its high thermodynamic reactivity and diradical state, its reactions are kinetically slow due to spin restriction. The class I ROS that include superoxide, hydrogen peroxide, or hydroxyl radicals can be produced when one or more electrons are transferred to O_2 (Fig. 6.1) (Rosner and Storz 1997).

Singlet oxygen ($^1O_2$) is a member of the class II ROS generated by energy transfer to molecular oxygen (also referred to as $^3O_2$) that removes the spin restriction and generates a highly reactive oxygen species. The energy transfer leads to rearrangement of the π*2p electrons and can produce two different forms of $^1O_2$: $^1\Delta g$ and $^1\Sigma g$. The $^1\Sigma g$ is extremely unstable and reverts to the longer-lived $^1\Delta g$ form. In $^1O_2$ ($^1\Delta g$; furthermore just referred to as $^1O_2$), one unpaired electron from a π*2p orbital in $^3O_2$ is excited and transferred to the other π* orbital (Fig. 6.1). This process creates a lone empty electron orbital, thus making $^1O_2$ a very powerful oxidant. The energy difference between $^3O_2$ and $^1O_2$ is 92 kJ M^{-1} or ~900 meV (Davies 2003; Nyman and Hynninen 2004; Grether-Beck et al. 2000; Steinberg 2012). Due to its high reactivity, $^1O_2$ has the ability to attack a variety of biomolecules, including proteins (Davies 2003, 2004, 2005; Clennan 2001), lipids (Glaeser and Klug 2005; Lupinkova and Komenda 2004; Nishiyama et al. 2004; Rinalducci et al. 2004), and nucleic acids (Sies 1993; Piette 1991).

The organization of outer orbital electrons in $^1O_2$, superoxide, hydrogen peroxide, and hydroxyl radicals imparts unique chemistry to each compound. The primary reaction products generated by $^1O_2$ differ from those generated by the class I ROS, and as a consequence, $^1O_2$ is not detoxified by activities that prevent the activity or repair the damages of class I ROS (Davies 2004, 2005; Clennan 2001; Nyman and Hynninen 2004). Analysis of $^1O_2$ reactivity with model compounds in vitro identified the thiol side chains of cysteine residues as prime targets for protein modification by this ROS

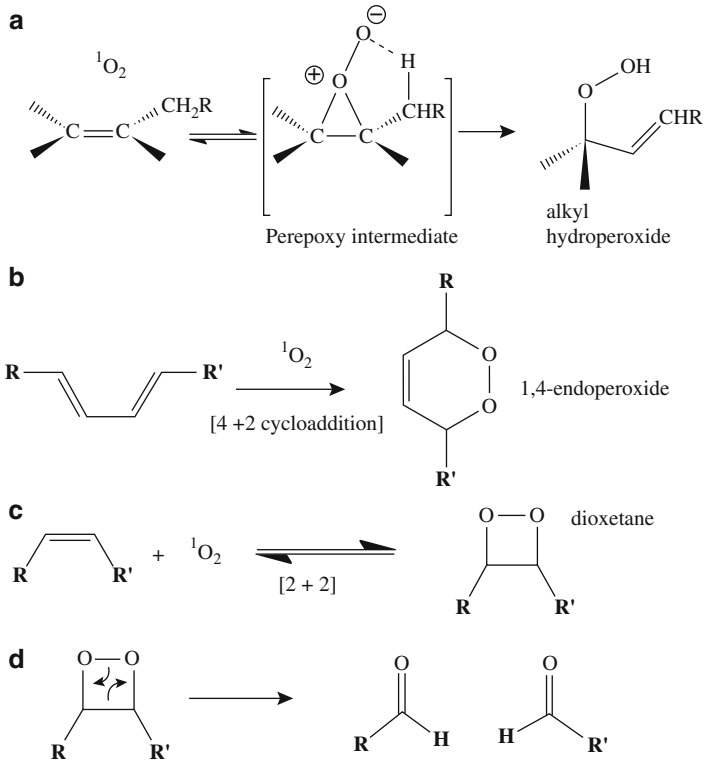

Fig. 6.2 Oxidation reaction products generated from singlet oxygen. Reaction products generated from oxidation by $^1O_2$ include (**a**) alkyl hydroperoxide, (**b**) 1,4-endoperoxide, and (**c**) dioxetane. In reaction (**a**), a perepoxy intermediate is formed before generation of the alkyl hydroperoxide, which subsequently can continue to react with nearby targets. The product of reaction (**b**) requires a conjugated double bond system to react with $^1O_2$ via [4+2] cycloaddition. The dioxetane product (**c**) can further generate two aldehyde compounds. Breakdown of the dioxetane product can destabilize membrane integrity when $^1O_2$ reacts at fatty acid sites of unsaturation (Landrum 2013)

(Clennan 2001), with additional studies showing that tryptophan, histidine, tyrosine, and methionine side chains may also be subject to attack (Estevam et al. 2004; Wright et al. 2002; Davies 2003, 2005). Attack of cysteine sulfhydryl groups by $^1O_2$ leads to the initial production of persulfoxides (Fig. 6.2) (Clennan 2001; Landrum 2013). Despite the importance and cellular toxicity of $^1O_2$, little is known about its interactions with biomolecules, particularly when compared to class I ROS.

The persulfoxide products, in turn, can attack neighboring functional groups (amino, aromatic, etc.) or bonds (C–C and C–N) and generate additional damage (Clennan 2001). Attack by $^1O_2$ on other amino acid side chains and on double bonds can generate endoperoxides (Fig. 6.2) (Davies 2003, 2004, 2005). These endoperoxides, like the persulfoxide intermediate, can continue to damage the protein and neighboring biomolecules, leading to amino acid side chain oxidation, peptide or

phosphodiester bond cleavage, or unsaturated fatty acid oxidation (Wright et al. 2000, 2002; Davies 2003, 2004; Kanofsky 1989a; Girotti and Kriska 2004). Due to both the reactivity of $^1O_2$ and cascade of oxidative damage that is triggered from $^1O_2$-mediated oxidation, it is not surprising then that prokaryotic and eukaryotic cells can succumb to the damage of this ROS.

Sources of $^1O_2$

Environmentally significant sources of $^1O_2$ exist, produced naturally or anthropogenically. The major process is energy distribution by a photosensitizer—a compound that, when in an energetically excited state, is capable of transferring that energy to molecular oxygen. Under photosynthetic growth conditions, capture of light energy by photosynthetic pigments (such as chlorophyll or bacteriochlorophyll) in the plant, algal or bacterial photosynthetic apparatus leads to excited-state pigments that can transfer that energy to molecular oxygen and generate $^1O_2$ (Anthony et al. 2005; Glaeser and Klug 2005; Lupinkova and Komenda 2004; Rinalducci et al. 2004; Glaeser et al. 2011; Ziegelhoffer and Donohue 2009).

Natural chromophores such as tetrapyrroles or porphyrin ring systems can act as photosensitizers to generate $^1O_2$ in the presence of light and molecular O_2. Light-independent reactions can also generate $^1O_2$, such that the catalytic activity of NADH oxidase and various peroxidases will produce $^1O_2$ as an inadvertent by-product in non-photosynthetic organisms (Davies 2005; Kanofsky 1983, 1984, 1988, 1989a, b, 1991; Kanofsky and Axelrod 1986; Kanofsky et al. 1988). In both animals and plants, the bactericidal activity of $^1O_2$ is used to defend against microbial pathogens (Davies 2004). Macrophages generate an oxidative burst as part of the host immune response to pathogens that includes formation of $^1O_2$ as well as other reactive oxygen and reactive nitrogen species (Kanofsky et al. 1988; Davies 2004; Kochevar 2004; Kanofsky 1991).

Chemically, $^1O_2$ can be generated by direct energy transfer from UV light to O_2, or by the action of light on a chemical photosensitizer like methylene blue or rose bengal (Davies 2004; Kochevar 2004). Photodynamic therapy relies on using chemical photosensitizers and a directed light source to generate $^1O_2$ and prevent cell growth, such as inhibiting the growth of cancer cells (Nyman and Hynninen 2004). An understanding of the mechanisms by which cells respond to and protect themselves from $^1O_2$ is of considerable interest for therapeutic, antimicrobial, and energetic reasons.

$^1O_2$ Production in Photosynthetic Organisms

It is believed that $^1O_2$ is the major damaging form of ROS in photosynthetic organisms (Triantaphylides et al. 2008). Since $^1O_2$ inhibits the productivity of the photosynthetic apparatus, a large amount of information is available regarding formation of

this ROS (Szabo et al. 2005; Cogdell et al. 2000; Frank and Brudvig 2004a; Kochevar 2004; Danon et al. 2005; Barber and Andersson 1992; Fryer et al. 2002). Light energy absorbed by chlorophyll (in plants) or bacteriochlorophyll (in prokaryotes) pigments found in the light-harvesting complexes of the photosynthetic apparatus leads to pigment excitation to a high-energy triplet state (Cogdell et al. 2000; Frank and Brudvig 2004a; Kochevar 2004). The excited pigments typically transferred this energy to the reaction center in order to generate reducing power in the form of reduced quinone molecules. In the presence of available molecular oxygen, the triplet state pigments can transfer energy to O_2 and generate $^1O_2$ (Cogdell et al. 2000; Frank and Brudvig 2004a; Kochevar 2004).

A major consequence of damage by $^1O_2$ in photosynthetic bacteria is the inactivation of bacterial reaction center complexes (Cogdell et al. 2000; Frank and Brudvig 2004a). In plants, major damage occurs to photosystem II (Lupinkova and Komenda 2004; Nishiyama et al. 2004) and $^1O_2$ can lead to the formation of necrotic lesions within the chloroplast (Danon et al. 2005). $^1O_2$ blocks repair of the light-sensitive D1 subunit of plant photosystem II (Rinalducci et al. 2004; Hideg et al. 2007; Kochevar 2004; Nishiyama et al. 2004), and leads to fatty acid peroxidation (Kriska and Girotti 2004; Girotti and Kriska 2004), which in turn causes destruction of the integrity of the lipid bilayer, abolishment of membrane enzyme activity (Kochevar 2004), and triggering of apoptosis (Danon et al. 2005; Foyer and Noctor 2005). Analysis of thylakoid membranes in plant cells in the absence of protective quenching compounds, such as carotenoids, shows increased levels of lipid peroxidation, a product of $^1O_2$-mediated oxidation (Triantaphylides et al. 2008).

There is no appreciable accumulation of $^1O_2$ in cells, in part due to its reactivity. Damage had been predicted to be localized, given the reactivity and proximity of likely target molecules to the site of $^1O_2$ generation. The cellular half-life of $^1O_2$ was initially estimated to be in ~100 ns range, implicating that singlet oxygen would only travel a short distance (<100 nm) before encountering and reacting with a target molecule (Davies 2004; Kochevar 2004). More recent studies have estimated the half-life of $^1O_2$ to be ~ 3 μs, increasing the potential diffusion area to be >250 nm and allowing $^1O_2$ to travel across a significant portion of the microbial cell and causing damage to a larger spectrum of biomolecules (Glaeser et al. 2011; Skovsen et al. 2005). The total global cellular damage caused by $^1O_2$ or the mechanisms by which cells recognize and respond to formation of any damage produced by this reactive oxygen species remains to be elucidated.

Energy Generation in *R. sphaeroides*

R. sphaeroides, a purple non-sulfur α-proteobacterium, is capable of growth in a wide variety of conditions and has been studied for decades as a model system for investigating anoxygenic photosynthesis (Blankenship et al. 1995). This facultative bacterium can grow under a diverse cadre of metabolic conditions, and is

normally found in freshwater or marine environments. In those environments, *R. sphaeroides* can be found growing photosynthetically or utilizing aerobic or anaerobic respiration (depending on the electron acceptors available). Additionally, under photosynthetic growth conditions *R. sphaeroides* is capable of autotrophic or heterotrophic growth. The regulation of genes encoding the photosynthetic machinery as well as genes involved in switching between growth types has been characterized to great detail.

More recently, *R. sphaeroides* has also become a model of economic importance for its use as a potential source of renewable bioproducts, such as biofuels and bioplastics. During photosynthetic growth, *R. sphaeroides* produces large amounts of carotenoids and isoprenoids that are of economic importance (Cogdell et al. 2000; Glaeser and Klug 2005; Glaeser et al. 2011; Monger et al. 1976; Niederman et al. 1976; Slouf et al. 2012). Polyhydroxybutyrate (PHB), a polymer of 3-hydroxybutyrate, is produced likely as a storage compound for carbon and reducing power storage but is of interest as a bioplastic material (Imam et al. 2011). When grown under nitrogen-limiting conditions, the activity of the nitrogenase enzyme complex in *R. sphaeroides* also leads to production of hydrogen (H_2) gas that can be exploited as a biofuel source (Kontur et al. 2011).

The generation of these bioproducts of interest is being coupling to the harnessing of energy generated from using light or solar energy, such that production is dependent on efficient energy utilization. Due to its metabolic diversity, one may expect *R. sphaeroides* to have evolved an array of regulatory mechanisms working in concert to ensure the most efficient processes for carbon utilization and/or energy generation or conservation based on the environmental conditions (Imam et al. 2011). These mechanisms are currently being investigated in order to exploit any potential bioproduct (Imam et al. 2011; Kontur et al. 2011). These investigations include the utilization of next-generation sequencing and multiple-omics (i.e., proteomics) technologies that allow the determination of global effects and can lead to maximum energy generation and/or compound production.

As previously stated, there is evidence that $^1O_2$ is the major damaging ROS in photosynthetic organisms (Triantaphylides et al. 2008). The implication is that $^1O_2$ would prevent maximum output of bioproducts of interest if the photosynthetic apparatus is damaged as a result of ROS production. As such, characterizing the mechanisms by which photosynthetic cells contend with $^1O_2$ is of importance. Substantial quantities of $^1O_2$ can be formed during photosynthetic growth of *R. sphaeroides* in the presence of molecular O_2. Additionally, $^1O_2$ can be generated by treatment of *R. sphaeroides* with the photosensitizing dye methylene blue when under aerobic conditions, limiting the production of carotenoids that are associated with photosynthetic growth (Anthony et al. 2005; Kumar et al. 2012; Nuss et al. 2010; Greenwell et al. 2011). The use of *R. sphaeroides* to characterize cellular responses to $^1O_2$ is of particular importance due to the years of research in photosynthesis, availability of systems capable of monitoring the effects of $^1O_2$ in vivo, and coupled with the use of genomic approaches make this bacterium the best model system (Ziegelhoffer and Donohue 2009).

Quenching of $^1O_2$ by Carotenoids

Carotenoids are a natural line of defense that can prevent damage by $^1O_2$ via quenching in both photosynthetic and non-photosynthetic organisms. This insight was initially discovered by studying the extreme sensitivity of carotenoid mutants of photosynthetic bacteria to light in the presence of oxygen (Sager and Zalokar 1958; Anderson and Robertson 1960; Griffiths et al. 1955). Carotenoids in photosynthetic microorganisms and in chloroplasts provide protection against singlet oxygen formation by quenching excited triplet-state chlorophylls and bacteriochlorophylls, or by directly quenching singlet directly (Cogdell et al. 2000; Krieger-Liszkay et al. 2008; Telfer 2005; Trebst 2003; Ziegelhoffer and Donohue 2009; Glaeser and Klug 2005). The energy transferred from excited chlorophylls or $^1O_2$ is transferred to and excites the carotenoid from its ground state. The excited carotenoid relaxes back to ground state by releasing the excess energy as heat (Domonkos et al. 2013; Frank and Brudvig 2004b; Johnson and Schroeder 1996; Telfer 2005; Young and Frank 1996). No chemical change to the carotenoid occurs during this process, leaving the molecule unchanged. A wide variety of carotenoid molecules exist, from β-carotene to α-tocopherol to xanthophylls, with well over 600 different carotenoids that have been studied (Kruk et al. 2005; Bendich and Olson 1989).

While carotenoids can function to directly quench $^1O_2$, the levels of this ROS can accumulate over time and overwhelm this defense, eventually leading to irreversible carotenoid modification (Ziegelhoffer and Donohue 2009). Since $^1O_2$ destroys proteins and lipids of the photosynthetic apparatus, the quenching activity of carotenoids must be insufficient to fully protect against this reactive oxygen species and as such, other systems must be utilized to protect and counteract $^1O_2$-mediated damage (Rinalducci et al. 2004; Kochevar 2004; Nishiyama et al. 2004; Hideg et al. 2000; Estevam et al. 2004; Ziegelhoffer and Donohue 2009).

Transcriptional Response to $^1O_2$ by *R. sphaeroides*

The activities of transcription factors are used to alter gene expression once an appropriate signal or stimulus is applied to the cell. These transcription factors vary in their structures, the mechanisms by which the signal is transduced into a response, and their interactions with either the DNA, RNA polymerase, or both. A specific transcriptional response to $^1O_2$ was identified in *R. sphaeroides*, and this response is essential when *R. sphaeroides* cells are limited in carotenoid production (Anthony et al. 2005). This transcriptional response is regulated by two proteins: the Group IV sigma (σ) factor, σ^E, and the anti-σ factor ChrR.

In bacteria, sigma (σ) factors function to regulate transcriptional activity by recruiting RNA polymerase (RNAP) to specific promoter elements. The number of σ factors encoded by different bacteria varies dramatically: the model bacterium *Escherichia coli* has only 7 σ factors total, whereas *Streptomyces coelicolor* has 64

different σ factors (~50 of which fall into the ECF sub-class of σ factors) (Bentley et al. 2002; Gross et al. 1996; Mascher 2013). The ECF σ factor family is the largest and most functionally diverse group of σ factors, and the transcriptional networks regulated by ECF σ factors allow cells to adapt and respond to a wide variety of environmental stimuli (Helmann 2002; Lonetto et al. 1992; Mascher 2013). Some common characteristics of ECF σ factors include positive auto-regulation of their own structural gene and co-transcription of their structural gene with a gene that encodes a cognate anti-σ factor that binds to and inhibits activity of the σ factor until the cell receives the appropriate environmental stimulus.

The activity of σ^E is controlled by the anti-sigma factor ChrR which interacts with σ^E in a heterodimeric complex in the absence of $^1O_2$ (Anthony et al. 2004; Campbell et al. 2007; Newman et al. 1999, 2001). Once bound to ChrR, σ^E is unable to bind core RNA polymerase or initiate transcription of target genes. ChrR does this by preventing the interaction of the σ factor with core RNA polymerase (Anthony et al. 2004). The anti-sigma factor ChrR is a zinc metalloprotein that requires this metal to inhibit σ^E activity (Newman et al. 2001).

It was shown that *R. sphaeroides* requires σ^E to transcribe genes needed for survival in the presence of $^1O_2$ when carotenoids are limiting, and that σ^E activity was not induced when cells were exposed to oxidants including superoxide, hydrogen peroxide, and hydroxyl radicals (Anthony et al. 2005). However, it has been shown that this system also responds to the *tertiary*-butyl hydroperoxide (*t*-BOOH), but it is presumed that this activation of σ^E through a different mechanism than induced by $^1O_2$ (Nam et al. 2013). The existing model is $^1O_2$ is sensed by ChrR, either directly or indirectly by an oxidation by-product, to subsequently release σ^E in order to initiate the transcriptional response (Greenwell et al. 2011; Nam et al. 2013). Initially, target genes regulated by σ^E were identified using gene expression studies (Anthony et al. 2005), whereas a more complete analysis others were identified using an in silico investigation utilizing a wide array of previously published microarray analyses and subsequently confirmed using chromatin immunoprecipitation on a chip (ChIP-chip) assays (Dufour et al. 2008).

Alternative Responses to $^1O_2$ by Other Organisms

While the σ^E–ChrR system of *R. sphaeroides* is not the only system triggered in response to $^1O_2$, to date it is the best characterized. A homolog of the σ^E–ChrR system in the non-photosynthetic bacterium *Caulobacter crescentus* has also been shown to be initiated in response to $^1O_2$, as well as *t*-BOOH, cadmium, and UV-A exposure (Lourenco and Gomes 2009).

Other organisms have different transcriptional responses initiated when exposed to $^1O_2$. The algae *Chlamydomonas reinhardti* has served as a eukaryotic model for monitoring the effects of exposure to $^1O_2$. The first evidence of specific gene induction due to $^1O_2$ was observed in *C. reinhardti* by monitoring expression of the *GPX5* gene that was robustly induced specifically to $^1O_2$ but

minimally affected by alternative ROS tested (Leisinger et al. 2001; Brzezowski et al. 2012; Fischer et al. 2009, 2010, 2012). Carotenoid biogenesis is induced in certain prokaryotes and eukaryotes. In *Myxococcus xanthus* it is induced by $^1O_2$ generated due to high levels of the tetrapyrrole protoporphyrin IX, which can act as a photosensitizer, accumulated during stationary phase (Botella et al. 1995; Browning et al. 2003; Galbis-Martinez et al. 2012; Martinez-Laborda et al. 1990; Martinez-Laborda and Murillo 1989; Whitworth et al. 2004). When grown in the presence of light, the obligate aerobe actinomycete *S. coelicolor* also induces carotenogenesis (Takano et al. 2005), grown under conditions that can generate $^1O_2$. This light-induced mechanism may be due to either detection of $^1O_2$ or as a preventive measure prior to $^1O_2$ generation.

As stated above, the response systems that contend with other ROS are considered to be not adapted for contending with $^1O_2$. However, some of those response systems are observed to be induced when cells encounter this reactive oxygen species. The SoxRS regulon of *E. coli* was induced in a *soxR*-dependent fashion by an exogenous endoperoxide compound, disodium 3,3′-(1,4-naphthylidene) diproprionate ($NDPO_2$) endoperoxides, that generates $^1O_2$ by decomposition (Agnez-Lima et al. 2001; Ziegelhoffer and Donohue 2009). Additionally, overexpression of OxyR, a regulator of hydrogen peroxide-inducible genes, in *E. coli* diminished the oxidative damage generated by $^1O_2$ (Schulz et al. 2000a; Agnez et al. 1996; Kim et al. 2002). One proposal for how these systems may protect cells from oxidative damage is due to the increased expression of antioxidant enzymes and scavenging compounds that may act to quench $^1O_2$ rather than repair any damage. Insights into the global effects of $^1O_2$, and any other stress or response mechanism, can be further investigated using more recently developed technologies.

Utilization of Next-Generation Sequencing (NGS) Technologies, Proteomic, and Metabolomic Approaches to Characterize Cellular Responses

Next-generation sequencing (NGS) technologies have been developing over the past decade, providing a wide variety of applications for improved investigations. For example, NGS can be utilized for variation analysis using whole-genome resequencing, transcriptome and non-coding RNA analyses via RNA sequencing (RNA-seq) technology, DNA–protein interactions via chromatin-immunoprecipitation with sequencing analysis (ChIP-seq), and a wide variety of other applications (Conway et al. 2012; Egan et al. 2012a, b). Combinatorial NGS technologies provide a breadth and depth of insight into research, allowing more comprehensive determination of effects on a more global scale than what has been attainable via reductive research investigations. In addition to the benefit of increased throughput via the use of NGS technologies, decreased cost of data generation has led to vast improvements in analyses—for example, resequencing of entire plant genomes is

no longer time- and cost-prohibitive to investigate genotype–phenotype relationships in complex systems (Rounsley and Last 2010). A number of NGS technologies and informational resources have become available very recently. Some commonly used platforms include the Roche 454 sequencing technology (Margulies et al. 2005), the Illumina Solexa Genome Analyzer (Bennett 2004; Bennett et al. 2005), and Life Technologies SOLiD or Ion Torrent platforms (Rothberg et al. 2011), to name a few.

As mentioned above, whole-genome sequencing using NGS technology allows for rapid and complete coverage of the genomes of bacteria, plants, and animals. Combinatorial approaches using multiple platforms in conjunction improves efficiency in assembly than a single platform, and sequencing alignment using a reference genome is obviously optimal. For example, several plant genomes have recently been sequenced by combinatorial NGS technologies and have allowed for the rapid acquisition of genome-scale variants data recognized by the high-throughput identification of mutations and alleles associated with various diverse phenotypes (DePristo et al. 2011).

For more complete investigation into transcriptomics above and beyond microarray analyses, RNA-seq is a particularly popular tool for the rapid collection and quantification of a much larger scale of RNAs both coding and non-coding (Wang et al. 2009, 2010; Garber et al. 2011). Coupling RNA-seq with other NGS technologies has allowed for simultaneous acquisition of genomic DNA sequence, profiles of gene expression, and detection of polymorphisms and splicing variants in order to derive resources and insight in a variety of plant species, including Arabidopsis and rice, that have particular economic and energetic importance (Filichkin et al. 2010; Gonzalez-Ballester et al. 2010; Castruita et al. 2011; Zenoni et al. 2010).

In addition to RNA-seq and other approaches that focus on transcriptional regulatory networks, other approaches include interactome analyses for networks formed via protein–protein interactions and metabolome analyses to monitor metabolic activity (Saito and Matsuda 2010). Metabolomic platforms couple ultra-performance liquid chromatography to mass spectrometry (Sawada et al. 2009). The accumulation and detection of a wide variety of metabolites in a large number of samples can occur rapidly and allow for the investigation of complex metabolic systems and changes in biological systems. Profiling of the metabolome provides a snapshot of metabolite accumulation in response to various conditions, including mutant analysis, treatments, and stress responses (Ishikawa et al. 2010; Kusano et al. 2011a, b, c). A recent metabolomic study looking at Arabidopsis mutants involved in methionine chain elongation showed that some of those enzymes were also involved in primary and secondary metabolite synthesis (Sawada et al. 2009).

Informational resources include the PRIMe (http://prime.psc.riken.jp/), MeRy-B (http://www.cbib.u-bordeaux2.fr/MERYB/), and MetabolomeExpress (https://www.metabolome-express.org/) that are particularly powerful analysis tools for metabolomic studies (Akiyama et al. 2008; Carroll et al. 2010; Ferry-Dumazet et al. 2011). In particular, a review by (Saito and Matsuda 2010) provides a summary list of resources available for metabolomic studies, particularly plant metabolomics.

Investigations into the Cellular Response to $^1O_2$ and Identification of the σ^E Regulon

The transcriptional response to $^1O_2$ identified in *R. sphaeroides* is regulated by the Group IV σ factor, σ^E, and the anti-σ factor ChrR (Anthony et al. 2005). In the absence of $^1O_2$, σ^E and ChrR interact in a heterodimeric complex preventing σ^E from binding core RNA polymerase to initiate transcription (Anthony et al. 2004; Campbell et al. 2007; Newman et al. 1999, 2001). It was previously shown that *R. sphaeroides* requires σ^E to transcribe genes needed for survival in the presence of $^1O_2$ when carotenoids are limiting, and that σ^E activity was not induced when cells were exposed to oxidants including superoxide, hydrogen peroxide, and hydroxyl radicals (Anthony et al. 2005). Some important characteristics of Group IV σ/anti-σ factor pairs include positive auto-regulation by the σ factor of its own coding gene, and cells lacking the anti-σ factor constitutively express the σ factor, and by virtue, the target genes regulated by the σ factor (Helmann 2002).

Using the characteristics describe for Group IV σ factor regulatory systems, the target genes regulated by σ^E were initially identified using gene expression studies via microarray, comparing gene expression in wild type *R. sphaeroides* to ΔChrR cells (Anthony et al. 2005). An in silico investigation that utilized previously published *R. sphaeroides* gene expression microarray analyses looked for genes that had correlated expression patterns with identified σ^E target genes via hierarchical clustering, and any genes identified as co-regulated were subsequently confirmed using ChIP-chip assays (Dufour et al. 2008).

In the initial microarray experiments, *rpoE* exhibited increased expression along with ~180 genes (corresponding to ~60 operons) that had a ≥3-fold increase in expression in ΔChrR cells when compared to wild type cells (Anthony et al. 2005). Only a small number of those operons were confirmed to be direct σ^E targets. Coupling this experiment to publicly available microarray analyses to identify co-regulated genes and CHiP-chip experiments to identify enriched DNA sequences occupied by σ^E in vivo allowed for a more complete determination of the net effect of $^1O_2$ on gene expression (Dufour et al. 2008, 2012). This work led to the identification of a core set of genes regulated by σ^E–ChrR in a wide variety of bacteria and an extended regulon that is found in α-proteobacteria (Dufour et al. 2008; Ziegelhoffer and Donohue 2009). Additional investigations into the response proposed that several small non-coding RNAs and the RNA chaperone Hfq also play a role in the cellular response to $^1O_2$, as Δhfq cells exhibited greater sensitivity and diminished induction of σ^E and members of the σ^E regulon when encountering $^1O_2$ (Berghoff et al. 2009, 2011).

Identified members of the core σ^E target regulon include *rpoEchrR*, the coding genes of the master regulators of this response to $^1O_2$ (Anthony et al. 2005; Dufour et al. 2008, 2012). The operon RSP2143–2144 is annotated to encode *phrA* and *cfaS*, respectively. The *phrA* gene encodes a DNA photolyase that can repair pyrimidine dimers formed by oxidative stress (Hendrischk et al. 2007; Dufour et al. 2008); *cfaS* is annotated as a cyclopropane fatty acid synthase predicted to modify membrane

lipids at sites of unsaturation by adding a methylene bridge, thus removing an oxidation target and maintaining the integrity of the fatty acid bilayer (Nam et al. 2013). The operon ranging from RSP1087–1091 contains uncharacterized gene products whose roles are unknown but are likely important to the cellular response, with the two genes encoded by RSP1091 and RSP1090 have limited amino acid similarity to cyclopropane fatty acid synthetases (Nam et al. 2013; Dufour et al. 2008; Ziegelhoffer and Donohue 2009). The *cycA* gene, encoded at locus tag RSP0296, is also a target gene regulated by and encodes cytochrome c_2 that functions as an electron shuttle in photosynthesis. Regulation of *cycA* is complex due to multiple transcription factors binding upstream and competing to regulate its expression, including the PrrA response regulator that regulates the shift between aerobic and photosynthetic growth, and the heat shock sigma factors $RpoH_I$ and $RpoH_{II}$ (Newman et al. 1999, 2001; Tavano et al. 2004; Eraso and Kaplan 1994; Karls et al. 1999). Other important functions are likely to be carried out by remaining members of the core σ^E–ChrR regulon but their functions are currently unknown (such as the RSP1087–1091 and the uncharacterized RSP1409 loci).

$RpoH_{II}$, one of two predicted homologs of the *E. coli* heat shock σ factor RpoH that are expressed in *R. sphaeroides*, in turn initiates transcription of a number of other genes that are involved in stress response. Inclusion of $RpoH_{II}$ comprises the extended σ^E–ChrR regulon found only in some α-proteobacteria, and thus σ^E activates a transcriptional cascade when it receives the inducing signal (Dufour et al. 2010). Further insight into the σ^E-dependent response and energy generation in *R. sphaeroides* requires globally based analytical techniques.

A multi "-omics" investigation of the global cellular dynamics of and response to exposure to $^1O_2$ was recently conducted (Berghoff et al. 2013). The cellular response to $^1O_2$ by *R. sphaeroides* was investigated on three levels: transcriptome analysis using microarrays and deep sequencing via RNA-seq, translatome analysis by polysome accumulation after treatment with chloramphenicol, and proteomic analysis via protein labeling using heavy amino acids coupled to liquid chromatography with tandem mass spectrometry (LC–MS/MS) (Berghoff et al. 2013). This investigation correlated transcriptional activity to translational activity of mRNAs as they are actively bound by ribosomes and to the actual levels of various proteins expressed in vivo. This is the first global investigation into the effects of $^1O_2$ on cellular transcription and translation activity and protein stability in vivo. These activities were compared over an extended exposure to $^1O_2$ at various time points to distinguish rapid, transient responses to long-lived induced or down-regulated activities. The dynamic alteration of activities included sulfur metabolic genes that were induced at the onset of stress but were not maintained for the duration of exposure (Berghoff et al. 2013). The classifications of activities altered globally included redox and stress defense mechanisms, and also included various metabolic activities such as carbohydrate, iron, and sulfur and amino acid metabolism, cell transport activities, and quorum sensing (Berghoff et al. 2013). This NGS and multiple-omics coupled investigation highlights the depth and ability to identify global effects of cellular responses, described above by the variety of activities affected by the presence of this ROS.

Coupling Insights to Improved Energy Generation

The use of newly developed NGS and proteomic analyses has great expended our ability to determine and understand the global ramifications of stresses and signals and the cellular responses to them. Specifically, the investigations described above have garnered significant insight into the global cellular effects of prolonged exposure to the ROS, $^1O_2$. Future investigations have to focus on identifying the functions and roles of uncharacterized gene products involved in the core and extended regulons of the *R. sphaeroides* σ^E–ChrR system as well as characterize the specific mechanisms by which $^1O_2$ alters the metabolic activities described above—are these alterations in metabolism simply due to damage caused directly by $^1O_2$ to essential components, are oxidative by-products such as organic hydroperoxides involved in a second level of oxidative damage that directly affects those cellular activities, or are $^1O_2$ and/or oxidative by-products recognized as signals to alter the activities such that there is minimal loss of efficient activity?

As $^1O_2$ is considered the major toxic compound that prohibits energy generation in photosynthetic centers, the σ^E–ChrR system can be used as part of a detection mechanism in *R. sphaeroides* against the presence of $^1O_2$. If $^1O_2$ is indeed the major ROS that affects energy generation, monitoring the response to the presence will be essential to ensure that maximum energy production is attained. Under growth conditions used to generate fuel source such as H_2, tracking σ^E activity during production will allow direct monitoring of a mechanism that threatens to decrease that energy potential. The use of this system as a monitor of energy potential in a fuel cell, where cellular exposure to molecular O_2 should be prohibited to maximize energy generation via photosynthetic growth conditions, provides an easily detectable assay.

While a significant amount of work remains to ascertain the mechanistic details described above, it also remains unclear as to whether the gene products or processes identified to combat $^1O_2$ are similar in prokaryotes and eukaryotes. Insights garnered from prokaryotic cellular responses to $^1O_2$ may lead to directed expression of genes necessary for cellular stability or energy production when cells or organisms are exposed to long durations of $^1O_2$. Since some of the genes induced by $^1O_2$ stress are proposed to confer protective or repair activities, other genes encode proteins of unknown function that currently prevent a comprehensive understanding of the cellular response to $^1O_2$. Since $^1O_2$ affects global cellular activities, the dynamic interplay and mechanisms by which those activities are altered also remain to be determined.

References

Agnez LF, Costa de Oliveira RL, Di Mascio P, Menck CF (1996) Involvement of *Escherichia coli* exonuclease III and endonuclease IV in the repair of singlet oxygen-induced DNA damage. Carcinogenesis 17(5):1183–1185

Agnez-Lima LF, Di Mascio P, Demple B, Menck CF (2001) Singlet molecular oxygen triggers the soxRS regulon of *Escherichia coli*. Biol Chem 382(7):1071–1075. doi:10.1515/BC.2001.134

Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T, Shinozaki K, Hirai MY, Sakurai T, Kikuchi J, Saito K (2008) PRIMe: a web site that assembles tools for metabolomics and transcriptomics. In Silico Biol 8(3–4):339–345

Anderson IC, Robertson DS (1960) Role of carotenoids in protecting chlorophyll from photodestruction. Plant Physiol 35(4):531–534

Anthony JR, Newman JD, Donohue TJ (2004) Interactions between the *Rhodobacter sphaeroides* ECF sigma factor, sigma(E), and its anti-sigma factor, ChrR. J Mol Biol 341(2):345–360. doi:10.1016/j.jmb.2004.06.018

Anthony JR, Warczak KL, Donohue TJ (2005) A transcriptional response to singlet oxygen, a toxic byproduct of photosynthesis. Proc Natl Acad Sci U S A 102(18):6502–6507. doi:10.1073/pnas.0502225102

Azzi A, Gennis RB (1986) Purification of the aa3-type cytochrome-c oxidase from *Rhodopseudomonas sphaeroides*. Methods Enzymol 126:138–145

Barber J, Andersson B (1992) Too much of a good thing: light can be bad for photosynthesis. Trends Biochem Sci 17 (2):61–66. pii: 0968-0004(92)90503-2

Bendich A, Olson JA (1989) Biological actions of carotenoids. FASEB J 3(8):1927–1932

Bennett S (2004) Solexa Ltd. Pharmacogenomics 5(4):433–438. doi:10.1517/14622416.5.4.433

Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005) Toward the 1,000 dollars human genome. Pharmacogenomics 6(4):373–382. doi:10.1517/14622416.6.4.373

Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417(6885):141–147. doi:10.1038/417141a

Berghoff BA, Glaeser J, Sharma CM, Vogel J, Klug G (2009) Photooxidative stress-induced and abundant small RNAs in *Rhodobacter sphaeroides*. Mol Microbiol 74(6):1497–1512. doi:10.1111/j.1365-2958.2009.06949.x

Berghoff BA, Glaeser J, Sharma CM, Zobawa M, Lottspeich F, Vogel J, Klug G (2011) Contribution of Hfq to photooxidative stress resistance and global regulation in *Rhodobacter sphaeroides*. Mol Microbiol 80(6):1479–1495. doi:10.1111/j.1365-2958.2011.07658.x

Berghoff BA, Konzer A, Mank NN, Looso M, Rische T, Forstner KU, Kruger M, Klug G (2013) Integrative "omics"-approach discovers dynamic and regulatory features of bacterial stress responses. PLoS Genet 9(6):e1003576. doi:10.1371/journal.pgen.1003576

Blankenship RE, Madigan MT, Bauer CE (eds) (1995) Anoxygenic photosynthetic bacteria. Kluwer Academic Press, Dordrecht

Botella JA, Murillo FJ, Ruiz-Vazquez R (1995) A cluster of structural and regulatory genes for light-induced carotenogenesis in *Myxococcus xanthus*. Eur J Biochem 233(1):238–248

Browning DF, Whitworth DE, Hodgson DA (2003) Light-induced carotenogenesis in *Myxococcus xanthus*: functional characterization of the ECF sigma factor CarQ and antisigma factor CarR. Mol Microbiol 48(1):237–251

Brzezowski P, Wilson KE, Gray GR (2012) The PSBP2 protein of *Chlamydomonas reinhardtii* is required for singlet oxygen-dependent signaling. Planta 236(4):1289–1303. doi:10.1007/s00425-012-1683-1

Campbell EA, Greenwell R, Anthony JR, Wang S, Lim L, Das K, Sofia HJ, Donohue TJ, Darst SA (2007) A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria. Mol Cell 27(5):793–805. doi:10.1016/j.molcel.2007.07.009

Carroll AJ, Badger MR, Harvey Millar A (2010) The MetabolomeExpress Project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. BMC Bioinformatics 11:376. doi:10.1186/1471-2105-11-376

Castruita M, Casero D, Karpowicz SJ, Kropat J, Vieler A, Hsieh SI, Yan W, Cokus S, Loo JA, Benning C, Pellegrini M, Merchant SS (2011) Systems biology approach in *Chlamydomonas* reveals connections between copper nutrition and multiple metabolic steps. Plant Cell 23(4):1273–1292. doi:10.1105/tpc.111.084400

Clennan EL (2001) Persulfoxide: key intermediate in reactions of singlet oxygen with sulfides. Acc Chem Res 34(11):875–884

Cogdell RJ, Howard TD, Bittl R, Schlodder E, Geisenheimer I, Lubitz W (2000) How carotenoids protect bacterial photosynthesis. Philos Trans R Soc Lond B Biol Sci 355(1402):1345–1349. doi:10.1098/rstb.2000.0696

Conway C, Chalkley R, High A, Maclennan K, Berri S, Chengot P, Alsop M, Egan P, Morgan J, Taylor GR, Chester J, Sen M, Rabbitts P, Wood HM (2012) Next-generation sequencing for simultaneous determination of human papillomavirus load, subtype, and associated genomic copy number changes in tumors. J Mol Diagn 14(2):104–111. doi:10.1016/j.jmoldx.2011.10.003

Danon A, Miersch O, Felix G, Camp RG, Apel K (2005) Concurrent activation of cell death-regulating signaling pathways by singlet oxygen in *Arabidopsis thaliana*. Plant J 41(1):68–80. doi:10.1111/j.1365-313X.2004.02276.x, pii: TPJ2276

Davies MJ (2003) Singlet oxygen-mediated damage to proteins and its consequences. Biochem Biophys Res Commun 305(3):761–770

Davies MJ (2004) Reactive species formed on proteins exposed to singlet oxygen. Photochem Photobiol Sci 3(1):17–25

Davies MJ (2005) The oxidative environment and protein damage. Biochim Biophys Acta 1703(2):93–109

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498. doi:10.1038/ng.806

Domonkos I, Kis M, Gombos Z, Ughy B (2013) Carotenoids, versatile components of oxygenic photosynthesis. Prog Lipid Res 52(4):539–561. doi:10.1016/j.plipres.2013.07.001

Dufour YS, Landick R, Donohue TJ (2008) Organization and evolution of the biological response to singlet oxygen stress. J Mol Biol 383(3):713–730. doi:10.1016/j.jmb.2008.08.017

Dufour YS, Kiley PJ, Donohue TJ (2010) Reconstruction of the core and extended regulons of global transcription factors. PLoS Genet 6(7):e1001027. doi:10.1371/journal.pgen.1001027

Dufour YS, Imam S, Koo BM, Green HA, Donohue TJ (2012) Convergence of the transcriptional responses to heat shock and singlet oxygen stresses. PLoS Genet 8(9):e1002929. doi:10.1371/journal.pgen.1002929

Egan AN, Schlueter J, Spooner DM (2012a) Applications of next-generation sequencing in plant biology. Am J Bot 99(2):175–185. doi:10.3732/ajb.1200020

Egan JB, Shi CX, Tembe W, Christoforides A, Kurdoglu A, Sinari S, Middha S, Asmann Y, Schmidt J, Braggio E, Keats JJ, Fonseca R, Bergsagel PL, Craig DW, Carpten JD, Stewart AK (2012b) Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. Blood 120(5):1060–1066. doi:10.1182/blood-2012-01-405977

Eraso JM, Kaplan S (1994) prrA, a putative response regulator involved in oxygen regulation of photosynthesis gene expression in *Rhodobacter sphaeroides*. J Bacteriol 176(1):32–43

Estevam ML, Nascimento OR, Baptista MS, Di Mascio P, Prado FM, Faljoni-Alario A, Zucchi Mdo R, Nantes IL (2004) Changes in the spin state and reactivity of cytochrome C induced by photochemically generated singlet oxygen and free radicals. J Biol Chem 279(38):39214–39222. doi:10.1074/jbc.M402093200, pii: M402093200

Ferry-Dumazet H, Gil L, Deborde C, Moing A, Bernillon S, Rolin D, Nikolski M, de Daruvar A, Jacob D (2011) MeRy-B: a web knowledgebase for the storage, visualization, analysis and annotation of plant NMR metabolomic profiles. BMC Plant Biol 11:104. doi:10.1186/1471-2229-11-104

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong WK, Mockler TC (2010) Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. Genome Res 20(1):45–58. doi:10.1101/gr.093302.109

Fischer BB, Dayer R, Schwarzenbach Y, Lemaire SD, Behra R, Liedtke A, Eggen RI (2009) Function and regulation of the glutathione peroxidase homologous gene GPXH/GPX5 in *Chlamydomonas reinhardtii*. Plant Mol Biol 71(6):569–583. doi:10.1007/s11103-009-9540-8

Fischer BB, Eggen RI, Niyogi KK (2010) Characterization of singlet oxygen-accumulating mutants isolated in a screen for altered oxidative stress response in *Chlamydomonas reinhardtii*. BMC Plant Biol 10:279. doi:10.1186/1471-2229-10-279

Fischer BB, Ledford HK, Wakao S, Huang SG, Casero D, Pellegrini M, Merchant SS, Koller A, Eggen RI, Niyogi KK (2012) SINGLET OXYGEN RESISTANT 1 links reactive electrophile signaling to singlet oxygen acclimation in *Chlamydomonas reinhardtii*. Proc Natl Acad Sci U S A 109(20):E1302–E1311. doi:10.1073/pnas.1116843109

Foyer CH, Noctor G (2005) Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses. Plant Cell 17(7):1866–1875. doi:10.1105/tpc.105.033589, pii: 17/7/1866

Frank HA, Brudvig GW (2004a) Redox functions of carotenoids in photosynthesis. Biochemistry 43:8605–8615

Frank HA, Brudvig GW (2004b) Redox functions of carotenoids in photosynthesis. Biochemistry 43(27):8607–8615. doi:10.1021/bi0492096

Fryer MJ, Oxborough K, Mullineaux PM, Baker NR (2002) Imaging of photo-oxidative stress responses in leaves. J Exp Bot 53(372):1249–1254

Galbis-Martinez M, Padmanabhan S, Murillo FJ, Elias-Arnanz M (2012) CarF mediates signaling by singlet oxygen, generated via photoexcited protoporphyrin IX, in *Myxococcus xanthus* light-induced carotenogenesis. J Bacteriol 194(6):1427–1436. doi:10.1128/JB.06662-11

Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8(6):469–477. doi:10.1038/nmeth.1613

Girotti AW, Kriska T (2004) Role of lipid hydroperoxides in photo-oxidative stress signaling. Antioxid Redox Signal 6(2):301–310. doi:10.1089/152308604322899369

Glaeser J, Klug G (2005) Photo-oxidative stress in *Rhodobacter sphaeroides*: protective role of carotenoids and expression of selected genes. Microbiology 151(Pt 6):1927–1938. doi:10.1099/mic.0.27789-0

Glaeser J, Nuss AM, Berghoff BA, Klug G (2011) Singlet oxygen stress in microorganisms. Adv Microb Physiol 58:141–173. doi:10.1016/B978-0-12-381043-4.00004-0

Gonzalez-Ballester D, Casero D, Cokus S, Pellegrini M, Merchant SS, Grossman AR (2010) RNA-seq analysis of sulfur-deprived *Chlamydomonas* cells reveals aspects of acclimation critical for cell survival. Plant Cell 22(6):2058–2084. doi:10.1105/tpc.109.071167

Greenwell R, Nam TW, Donohue TJ (2011) Features of *Rhodobacter sphaeroides* ChrR required for stimuli to promote the dissociation of sigma(E)/ChrR complexes. J Mol Biol 407(4):477–491. doi:10.1016/j.jmb.2011.01.055

Grether-Beck S, Bonizzi G, Schmitt-Brenden H, Felsner I, Timmer A, Sies H, Johnson JP, Piette J, Krutmann J (2000) Non-enzymatic triggering of the ceramide signalling cascade by solar UVA radiation. EMBO J 19(21):5793–5800

Griffiths M, Sistrom WR, Cohenbazire G, Stanier RY, Calvin M (1955) Function of carotenoids in photosynthesis. Nature 176(4495):1211–1215

Gross CA, Chan CL, Lonetto MA (1996) A structure/function analysis of *Escherichia coli* RNA polymerase. Philos Trans R Soc Lond B Biol Sci 351(1339):475–482. doi:10.1098/rstb.1996.0045

Helmann JD (2002) The extracytoplasmic function (ECF) sigma factors. Adv Microb Physiol 46:47–110

Hendrischk AK, Braatsch S, Glaeser J, Klug G (2007) The phrA gene of *Rhodobacter sphaeroides* encodes a photolyase and is regulated by singlet oxygen and peroxide in a sigma(E)-dependent manner. Microbiology 153(Pt 6):1842–1851. doi:10.1099/mic.0.2006/004390-0

Hideg E, Kalai T, Hideg K, Vass I (2000) Do oxidative stress conditions impairing photosynthesis in the light manifest as photoinhibition? Philos Trans R Soc Lond B Biol Sci 355(1402):1511–1516. doi:10.1098/rstb.2000.0711

Hideg E, Kos PB, Vass I (2007) Photosystem II damage induced by chemically generated singlet oxygen in tobacco leaves. Physiol Plant 131(1):33–40. doi:10.1111/j.1399-3054.2007.00913.x, pii: PPL913

Imam S, Yilmaz S, Sohmen U, Gorzalski AS, Reed JL, Noguera DR, Donohue TJ (2011) iRsp1095: a genome-scale reconstruction of the *Rhodobacter sphaeroides* metabolic network. BMC Syst Biol 5:116. doi:10.1186/1752-0509-5-116

Ishikawa T, Takahara K, Hirabayashi T, Matsumura H, Fujisawa S, Terauchi R, Uchimiya H, Kawai-Yamada M (2010) Metabolome analysis of response to oxidative stress in rice suspension cells overexpressing cell death suppressor Bax inhibitor-1. Plant Cell Physiol 51(1):9–20. doi:10.1093/pcp/pcp162

Johnson EA, Schroeder WA (1996) Microbial carotenoids. Adv Biochem Eng Biotechnol 53:119–178

Kanofsky JR (1983) Singlet oxygen production by lactoperoxidase. J Biol Chem 258(10):5991–5993

Kanofsky JR (1984) Singlet oxygen production by chloroperoxidase–hydrogen peroxide–halide systems. J Biol Chem 259(9):5596–5600

Kanofsky JR (1988) Singlet oxygen production from the peroxidase-catalyzed oxidation of indole-3-acetic acid. J Biol Chem 263(28):14171–14175

Kanofsky JR (1989a) Singlet oxygen production by biological systems. Chem Biol Interact 70(1–2):1–28

Kanofsky JR (1989b) Singlet oxygen production from the peroxidase catalyzed formation of styrene glutathione adducts. Biochem Biophys Res Commun 159(3):1051–1054

Kanofsky JR (1991) Quenching of singlet oxygen by human red cell ghosts. Photochem Photobiol 53(1):93–99

Kanofsky JR, Axelrod B (1986) Singlet oxygen production by soybean lipoxygenase isozymes. J Biol Chem 261(3):1099–1104

Kanofsky JR, Hoogland H, Wever R, Weiss SJ (1988) Singlet oxygen production by human eosinophils. J Biol Chem 263(20):9692–9696

Karls RK, Wolf JR, Donohue TJ (1999) Activation of the cycA P2 promoter for the *Rhodobacter sphaeroides* cytochrome c2 gene by the photosynthesis response regulator. Mol Microbiol 34(4):822–835

Kerr RA (2005) Earth science. The story of O_2. Science 308(5729):1730–1732. doi:10.1126/science.308.5729.1730

Kiley PJ, Storz G (2004) Exploiting thiol modifications. PLoS Biol 2(11):e400. doi:10.1371/journal.pbio.0020400

Kim SY, Kim EJ, Park JW (2002) Control of singlet oxygen-induced oxidative damage in *Escherichia coli*. J Biochem Mol Biol 35(4):353–357

Kochevar I (2004) Singlet oxygen signaling: from intimate to global. Sci STKE. doi:10.1126/stke.2212004pe7

Kontur WS, Ziegelhoffer EC, Spero MA, Imam S, Noguera DR, Donohue TJ (2011) Pathways involved in reductant distribution during photobiological H(2) production by *Rhodobacter sphaeroides*. Appl Environ Microbiol 77(20):7425–7429. doi:10.1128/AEM.05273-11

Krieger-Liszkay A, Fufezan C, Trebst A (2008) Singlet oxygen production in photosystem II and related protection mechanism. Photosynth Res 98(1–3):551–564. doi:10.1007/s11120-008-9349-3

Kriska T, Girotti AW (2004) Separation and quantitation of peroxidized phospholipids using high-performance thin-layer chromatography with tetramethyl-*p*-phenylenediamine detection. Anal Biochem 327(1):97–106. doi:10.1016/j.ab.2003.12.021, pii: S0003269704000090

Kruk J, Hollander-Czytko H, Oettmeier W, Trebst A (2005) Tocopherol as singlet oxygen scavenger in photosystem II. J Plant Physiol 162(7):749–757. doi:10.1016/j.jplph.2005.04.020

Kumar S, Rai AK, Mishra MN, Shukla M, Singh PK, Tripathi AK (2012) RpoH2 sigma factor controls the photooxidative stress response in a non-photosynthetic rhizobacterium, *Azospirillum brasilense* Sp7. Microbiology 158(Pt 12):2891–2902. doi:10.1099/mic.0.062380-0

Kusano M, Jonsson P, Fukushima A, Gullberg J, Sjostrom M, Trygg J, Moritz T (2011a) Metabolite signature during short-day induced growth cessation in populus. Front Plant Sci 2:29. doi:10.3389/fpls.2011.00029

Kusano M, Tabuchi M, Fukushima A, Funayama K, Diaz C, Kobayashi M, Hayashi N, Tsuchiya YN, Takahashi H, Kamata A, Yamaya T, Saito K (2011b) Metabolomics data reveal a crucial role of cytosolic glutamine synthetase 1;1 in coordinating metabolic balance in rice. Plant J 66(3):456–466. doi:10.1111/j.1365-313X.2011.04506.x

Kusano M, Tohge T, Fukushima A, Kobayashi M, Hayashi N, Otsuki H, Kondou Y, Goto H, Kawashima M, Matsuda F, Niida R, Matsui M, Saito K, Fernie AR (2011c) Metabolomics reveals comprehensive reprogramming involving two independent metabolic responses of Arabidopsis to UV-B light. Plant J 67(2):354–369. doi:10.1111/j.1365-313X.2011.04599.x

Landrum JT (2013) Reactive oxygen and nitrogen species in biological systems: reactions and regulation by carotenoids. In: Tanumihardjo SA (ed) Carotenoids and human health. Nutrition and health. Humana Press, pp 57–101. doi:10.1007/978-1-62703-203-2_4

Leisinger U, Rufenacht K, Fischer B, Pesaro M, Spengler A, Zehnder AJ, Eggen RI (2001) The glutathione peroxidase homologous gene from *Chlamydomonas reinhardtii* is transcriptionally up-regulated by singlet oxygen. Plant Mol Biol 46(4):395–408

Lonetto M, Gribskov M, Gross CA (1992) The sigma 70 family: sequence conservation and evolutionary relationships. J Bacteriol 174(12):3843–3849

Lourenco RF, Gomes SL (2009) The transcriptional response to cadmium, organic hydroperoxide, singlet oxygen and UV-A mediated by the sigmaE-ChrR system in *Caulobacter crescentus*. Mol Microbiol 72(5):1159–1170. doi:10.1111/j.1365-2958.2009.06714.x

Lupinkova L, Komenda J (2004) Oxidative modifications of the Photosystem II D1 protein by reactive oxygen species: from isolated protein to cyanobacterial cells. Photochem Photobiol 79(2):152–162

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380. doi:10.1038/nature03959

Martinez-Laborda A, Murillo FJ (1989) Genic and allelic interactions in the carotenogenic response of *Myxococcus xanthus* to blue light. Genetics 122(3):481–490

Martinez-Laborda A, Balsalobre JM, Fontes M, Murillo FJ (1990) Accumulation of carotenoids in structural and regulatory mutants of the bacterium *Myxococcus xanthus*. Mol Gen Genet 223(2):205–210

Mascher T (2013) Signaling diversity and evolution of extracytoplasmic function (ECF) sigma factors. Curr Opin Microbiol 16(2):148–155. doi:10.1016/j.mib.2013.02.001

Miller MJ, Gennis RB (1986) Purification and reconstitution of the cytochrome d terminal oxidase complex from *Escherichia coli*. Methods Enzymol 126:87–94

Monger TG, Cogdell RJ, Parson WW (1976) Triplet states of bacteriochlorophyll and carotenoids in chromatophores of photosynthetic bacteria. Biochim Biophys Acta 449(1):136–153

Nam TW, Ziegelhoffer EC, Lemke RA, Donohue TJ (2013) Proteins needed to activate a transcriptional response to the reactive oxygen species singlet oxygen. mBio 4(1):e00541-12. doi:10.1128/mBio.00541-12

Newman JD, Falkowski MJ, Schilke BA, Anthony LC, Donohue TJ (1999) The *Rhodobacter sphaeroides* ECF sigma factor, sigma(E), and the target promoters cycA P3 and rpoE P1. J Mol Biol 294(2):307–320. doi:10.1006/jmbi.1999.3263

Newman JD, Anthony JR, Donohue TJ (2001) The importance of zinc-binding to the function of *Rhodobacter sphaeroides* ChrR as an anti-sigma factor. J Mol Biol 313(3):485–499. doi:10.1006/jmbi.2001.5069

Niederman RA, Mallon DE, Langan JJ (1976) Membranes of *Rhodopseudomonas sphaeroides*. IV. Assembly of chromatophores in low-aeration cell suspensions. Biochim Biophys Acta 440(2):429–447

Nishiyama Y, Allakhverdiev SI, Yamamoto H, Hayashi H, Murata N (2004) Singlet oxygen inhibits the repair of photosystem II by suppressing the translation elongation of the D1 protein in *Synechocystis* sp. PCC 6803. Biochemistry 43(35):11321–11330

Nuss AM, Glaeser J, Berghoff BA, Klug G (2010) Overlapping alternative sigma factor regulons in the response to singlet oxygen in *Rhodobacter sphaeroides*. J Bacteriol 192(10):2613–2623. doi:10.1128/JB.01605-09

Nyman ES, Hynninen PH (2004) Research advances in the use of tetrapyrrolic photosensitizers for photodynamic therapy. J Photochem Photobiol B 73(1–2):1–28

Piette J (1991) Biological consequences associated with DNA oxidation mediated by singlet oxygen. J Photochem Photobiol B 11(3–4):241–260

Rinalducci S, Pedersen JZ, Zolla L (2004) Formation of radicals from singlet oxygen produced during photoinhibition of isolated light-harvesting proteins of photosystem II. Biochim Biophys Acta 1608(1):63–73

Rosner JL, Storz G (1997) Regulation of bacterial responses to oxidative stress. Curr Top Cell Regul 35:163–177

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352. doi:10.1038/nature10242

Rounsley SD, Last RL (2010) Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. Plant J 61(6):922–927. doi:10.1111/j.1365-313X.2009.04030.x

Sager R, Zalokar M (1958) Pigments and photosynthesis in a carotenoid-deficient mutant of *Chlamydomonas*. Nature 182(4628):98–100

Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489. doi:10.1146/annurev.arplant.043008.092035

Sawada Y, Akiyama K, Sakata A, Kuwahara A, Otsuki H, Sakurai T, Saito K, Hirai MY (2009) Widely targeted metabolomics based on large-scale MS/MS data for elucidating metabolite accumulation patterns in plants. Plant Cell Physiol 50(1):37–47. doi:10.1093/pcp/pcn183

Schulz I, Mahler HC, Boiteux S, Epe B (2000a) Oxidative DNA base damage induced by singlet oxygen and photosensitization: recognition by repair endonucleases and mutagenicity. Mutat Res 461(2):145–156

Schulz JB, Dehmer T, Schols L, Mende H, Hardt C, Vorgerd M, Burk K, Matson W, Dichgans J, Beal MF, Bogdanov MB (2000b) Oxidative stress in patients with Friedreich ataxia. Neurology 55(11):1719–1721

Schulz JB, Lindenau J, Seyfried J, Dichgans J (2000c) Glutathione, oxidative stress and neurodegeneration. Eur J Biochem 267(16):4904–4911

Sies H (1993) Damage to plasmid DNA by singlet oxygen and its protection. Mutat Res 299(3–4):183–191

Skovsen E, Snyder JW, Lambert JD, Ogilby PR (2005) Lifetime and diffusion of singlet oxygen in a cell. J Phys Chem B 109(18):8570–8573. doi:10.1021/jp051163i

Slouf V, Chabera P, Olsen JD, Martin EC, Qian P, Hunter CN, Polivka T (2012) Photoprotection in a purple phototrophic bacterium mediated by oxygen-dependent alteration of carotenoid excited-state properties. Proc Natl Acad Sci U S A 109(22):8570–8575. doi:10.1073/pnas.1201413109

Steinberg CEW (2012) Activation of oxygen: multipurpose tool. In: Stress ecology. Springer, Netherlands, pp 7–45. doi:10.1007/978-94-007-2072-5_2

Storz G, Imlay JA (1999) Oxidative stress. Curr Opin Microbiol 2(2):188–194

Szabo I, Bergantino E, Giacometti GM (2005) Light and oxygenic photosynthesis: energy dissipation as a protection mechanism against photo-oxidation. EMBO Rep 6(7):629–634. doi:10.1038/sj.embor.7400460, pii: 7400460

Takano H, Obitsu S, Beppu T, Ueda K (2005) Light-induced carotenogenesis in *Streptomyces coelicolor* A3(2): identification of an extracytoplasmic function sigma factor that directs photodependent transcription of the carotenoid biosynthesis gene cluster. J Bacteriol 187(5):1825–1832. doi:10.1128/JB.187.5.1825-1832.2005

Tavano CL, Comolli JC, Donohue TJ (2004) The role of dor gene products in controlling the P2 promoter of the cytochrome c2 gene, cycA, in *Rhodobacter sphaeroides*. Microbiology 150(Pt 6):1893–1899. doi:10.1099/mic.0.26971-0

Telfer A (2005) Too much light? How beta-carotene protects the photosystem II reaction centre. Photochem Photobiol Sci 4(12):950–956. doi:10.1039/b507888c

Trebst A (2003) Function of beta-carotene and tocopherol in photosystem II. Z Naturforsch C 58(9–10):609–620

Triantaphylides C, Krischke M, Hoeberichts FA, Ksas B, Gresser G, Havaux M, Van Breusegem F, Mueller MJ (2008) Singlet oxygen is the major reactive oxygen species involved in photooxidative damage to plants. Plant Physiol 148(2):960–968. doi:10.1104/pp. 108.125690

Wang Z, Gerstein M, Snyder M (2009) RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10(1):57–63. doi:10.1038/nrg2484

Wang L, Feng Z, Wang X, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics 26(1):136–138. doi:10.1093/bioinformatics/btp612

Whitworth DE, Bryan SJ, Berry AE, McGowan SJ, Hodgson DA (2004) Genetic dissection of the light-inducible carQRS promoter region of *Myxococcus xanthus*. J Bacteriol 186(23):7836–7846. doi:10.1128/JB.186.23.7836-7846.2004

Wright A, Hawkins CL, Davies MJ (2000) Singlet oxygen-mediated protein oxidation: evidence for the formation of reactive peroxides. Redox Rep 5(2–3):159–161

Wright A, Bubb WA, Hawkins CL, Davies MJ (2002) Singlet oxygen-mediated protein oxidation: evidence for the formation of reactive side chain peroxides on tyrosine residues. Photochem Photobiol 76(1):35–46

Young AJ, Frank HA (1996) Energy transfer reactions involving carotenoids: quenching of chlorophyll fluorescence. J Photochem Photobiol B 36(1):3–15. doi:10.1016/S1011-1344(96)07397-6

Zenoni S, Ferrarini A, Giacomelli E, Xumerle L, Fasoli M, Malerba G, Bellin D, Pezzotti M, Delledonne M (2010) Characterization of transcriptional complexity during berry development in *Vitis vinifera* using RNA-seq. Plant Physiol 152(4):1787–1795. doi:10.1104/pp.109.149716

Zheng M, Storz G (2000) Redox sensing by prokaryotic transcription factors. Biochem Pharmacol 59(1):1–6

Ziegelhoffer EC, Donohue TJ (2009) Bacterial responses to photo-oxidative stress. Nat Rev Microbiol 7(12):856–863. doi:10.1038/nrmicro2237

Chapter 7
Experimental Evolution and Next Generation Sequencing Illuminate the Evolutionary Trajectories of Microbes

Mario A. Fares

Introduction

In his book "The origin of species by means of natural selection" (Darwin 1859), Charles Darwin manifested a deep frustration justified by the realization that natural selection is too slow to be observed in real time. He admittedly based all his conclusions on observations or indirect measurements of the action of natural selection and reported many evidence supporting his conclusion: "That natural selection will always act with extreme slowness I fully admit."

Darwin, if lived today, would be enthralled by the fact that the process of natural selection and the mechanisms underlying them could be directly tested in a reasonable short time using microbes. Microbes offer a unique opportunity to observe and test the mechanism of natural selection and the general principles of evolution. This is mainly due to the short generation times, small genome sizes, and deep microbes genetic and physiological characterization. These features and the feasibility of evolving microbes in the laboratory with the current technology under controlled conditions and at high "speeds" make them ideal systems to put the main principles of evolution to test and unearth the dynamics underlying the evolution of biological complexity (Kawecki et al. 2012). In addition to the possibility of conducting laboratory-supervised evolution experiments, the next generation sequencing technology (NGS) has enabled sequencing hundreds of microbial genomes at once, linking particular genome dynamics to microbes' lifestyles.

M.A. Fares (✉)
Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas (CSIC-UPV), Ingeniero Fausto Elio, Valencia 46022, Spain

Department of Genetics, Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin, Ireland
e-mail: mfares@ibmcp.upv.es, faresm@tcd.ie

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_7

In this chapter, I will discuss the many different scenarios under which microbes have been evolved in the laboratory, how did NGS contribute to the understanding of the genomes dynamics behind specific adaptive processes, and the main conceptual breakthroughs derived from these studies.

What Makes Microbes Attractive to Test Evolutionary Processes?

Eighty-five years ago, August Krogh articulated a principle (Krogh principle) after which experimentalists should choose the model organism that can best foster a clear and direct experimental design and a rigorous and unambiguous result and interpretation (Krogh 1929). Krogh principle is particularly useful when testing evolutionary processes, as these are often dominated by very complex patterns that are intermingled and many times shaped by the environment.

The general principles of Evolutionary biology has been historically built based on indirect theoretical and comparative studies (Futuyma 1998), lacking rigorous experimentation proof. There are several reasons for the lack of experimental studies probing principles of evolution. Mainly, it remains difficult identifying the dynamics of natural selection leading to the fixation of advantageous mutations at specific episode of organisms' evolution. Some of the reasons for this difficulty are the impracticality of replicating the complex mix of environmental conditions under which populations grew at some stage during their evolution and the slow pace at which natural selection acts. In this sense, microorganisms offer a unique opportunity for studying evolution as they present large populations sizes, short generation times, small genome sizes, and enormous physiological plasticity. Noticeably, microorganisms are not equipped with complex homeostasis systems, and thus their phenotype is largely the result of their genetic composition interacting with the environment. In addition to this convenient feature, microbes present a puzzling diversification whether measured in terms of the number of species (Dykhuizen 1998; Gans et al. 2005), habitat range (Pikuta et al. 2007), or the breadth of energy sources and biochemical pathways they can exploit in order to survive (Pace 1997).

The hallmarks of experimentation of any kind are control and replication. In evolutionary biology, controlling environmental conditions, especially when conducting experiments out of the laboratory, is difficult if not impossible. However, the fact that enormous population size of microbes could grow in tiny spaces (for example, a drop of culture medium) makes it feasible growing hundreds of microbial populations in a standard laboratory space. Moreover, microbiologists have successfully harnessed bacterial evolution and domesticated them to grow under laboratory-controlled conditions. Hundreds of microbial populations can be then propagated and analyzed simultaneously. If maintained evolving separately, with no cross-contamination, such populations can be used to test the repeatability of evolutionary processes (Lenski et al. 2000), to understand the physiological plasticity of bacteria growing under different carbon sources, and reproduce ecological scenarios of more complex organisms. In summary, experimental evolution allows determining

the selective forces operating, and by virtue of replicating the experiment, researchers can distinguish between deterministic and stochastic effects.

Environmental control is one of the most important advantages of using microbial populations because we can grow homogenously distributed populations in an environment in which single factors can be modified. In this new single-factor modified environments, that reproduces ancestral environments, many hypotheses can be tested, including how novel physiologies emerge to adapt to a new environment, the population dynamics of generalists and specialists, and the role of contingency in the adaptation to novel conditions and the trade-offs that such adaptations involve (Bennett and Lenski 2007; Bronikowski et al. 2001; Lee et al. 2009).

The large population sizes of microbes offers an analytical advantage, which is concerned with the higher likelihood of originating novel adaptations through mutations. The rationale is simple: in a small space of culture liquid billions of microbial cells can be kept and propagated, thereby avoiding the effect of genetic drift and directly testing the role of natural selection. During DNA replication, or even protein translation, there is a low but finite probability of an error in replication. The probability of occurrence of such a mutation is the product of the population size and the mutation rate. Therefore, the larger the population size the greater is the number of mutations originating in the population and the higher is the probability of a mutational novelty emerging. Because selection is strong when population sizes are large, the probability of fixation of beneficial mutations is very high. It follows then that the rate at which evolution occurs is high in microbial populations, making it possible reproducing adaptive evolution in real time. Indeed, in long-term evolutionary studies on microbial populations, every single nucleotide base pair should have experienced at least one mutation, and thus have undergone selection filtering (Lenski et al. 2003).

Finally, unlike multicellular organisms that require at least days or weeks to produce a new generation, microbes require minutes or hours. This allows beneficial mutations to become quickly fixed in the populations. For example, thermo-resistant mutations can become fixed in the microbial population within 15–20 days after initiating an evolution experiment (Bennett and Lenski 2007; Elena and Lenski 2003). Moreover, the enormous linkage disequilibrium of microbes ensures their clonal transmission for thousands of generations preserving the ancestral genetic background. This, in addition to the possibility of freezing evolved cells that can be thawed again, permits building a microbial fossil record and perform genome archeology at any time of the evolution experiment (Lenski et al. 2003; Ostrowski et al. 2008).

Experimental Evolution and Mutation Accumulation Dynamics

Experimental evolution combined with whole-genome re-sequencing is a promising strategy for investigating the dynamics of evolutionary change. One of the questions that have motivated efforts in reproducing an evolutionary scenario is how repeatable is evolution. The fragmentary nature of the fossil record cannot provide a full

picture that would allow answering this question, and even if it did we are not certain what kinds of environments or adaptations have not been explored by nature. Instead, reproducing fine-tuned scenarios in a test tube containing billions of bacterial cells can shed light on the complexity of evolutionary patterns.

Evolution experiments start with an initial population of microbes genetically identical and adapted to an ancestral environment (Fig. 7.1). Adaptation is determined by the Malthusian growth parameter of the population and is considered to be proportional to the relative fitness of the population. Fitness in experimentally evolved populations is measured as the capacity of such descendent populations to compete head-to-head with their ancestors. These two populations, the evolved one and its parental ancestral population, can be compared because they can be brought together in the same place at the same time. We can compare the performance of the descendant and ancestral populations by quantifying the number of offspring that each leaves in the next generations in an environment in which the carbon source is common for the two differentiated populations. Populations are propagated between

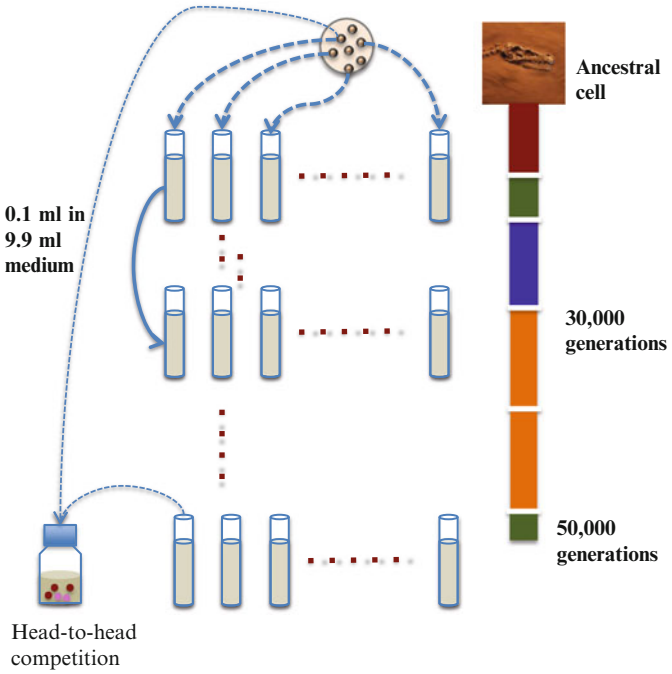

Fig. 7.1 Experimental evolution of microbes in the laboratory. From a single cell (ancestral cell) many replicate populations are generated, all being genetically identical and evolving for many generations independently. To assay the biological fitness of the evolved population at any time point, equal proportions of this population and of its ancestral one are mixed in the same medium. Both cells, the evolved and ancestral ones, should be distinguishable, for example through a metabolic marker that yields distinctly colored cells, to determine the relative frequency of each population at the start and end of the competition experiment. Improved fitness of the evolved population is reflected in a higher proportion of evolved cells than ancestors in the competition experiment

generations by diluting 0.1 ml of the grown culture in 9.9 ml of a new culture. To determine how repeatable is evolution, many different independently evolving lineages are generated from the same ancestor, and thus originally presenting the same genetic background and evolved in parallel (Fig. 7.1). The many different evolutionary paths followed by each of the independent evolving lines can be then compared and their differences quantified.

As I explained earlier, microbes are genetically represented by one chromosome. The gamma-proteobacterium *Escherichia coli* strain K12 MG1655 is the one most used in experimental evolution of microbes. Most bacteria, including *E. coli*, present highly dense genomes, with the genome size reflecting the number of genes (Giovannoni et al. 2005; Mira et al. 2001). The high gene density of these genomes and large linkage disequilibrium means that the mutational load is expected to increase as generations pass by without disrupting previous genetic backgrounds and that most changes will be affecting coding genes or regulatory regions. This means that we can directly associate particular nucleotide mutations to specific phenotypes and follow the history of interesting mutations since the last common ancestor of all the founded bacterial populations. Likewise, the yeast *Saccharomyces cerevisiae* has been used in its haploid or diploid genetic structure as a model to test specific evolutionary processes through experimental evolution. Here I provide examples of how NGS performs a powerful tool when combined with experimental evolution to unearth the rules governing fundamental evolutionary processes.

The Evolutionary Trajectories of Adaptive Mutations

NGS has been developed reaching a stage in which single minority mutations can be identified at low frequencies and their origin traced through reviving evolved cells at different time points of an evolution experiment. For example, the final stages of the fixation of an adaptive mutation can be identified by mixing equal proportions of bacterial cells labeled with two different tags (Hegreness et al. 2006). Combining cost-effective Illumina re-sequencing with experimental evolution makes it possible to sequence several hundreds of individuals from an evolved population, generating estimates of allele frequencies at millions of single-nucleotide polymorphisms (SNPs) genome-wide (Burke 2012; Burke et al. 2010; Burke and Long 2012; Futschik and Schlotterer 2010). This is important not only to identify rare variants but also to determine with unprecedented accuracy the evolutionary trajectories of adaptive mutations.

Evolution experiments seeking to identify adaptive evolution derive populations from a single ancestral genotypes, and thus genetically identical, in a constant environment or an environment with constant fluctuations. This is achieved by a continuous culture of populations in which the input of resources and the removal of individuals occur at a constant and controlled way. Alternatively, a fraction of the grown population is passaged to a new culture medium. When an adaptive mutation emerges in such an environment, this drives the evolutionary dynamic of the population, so that the

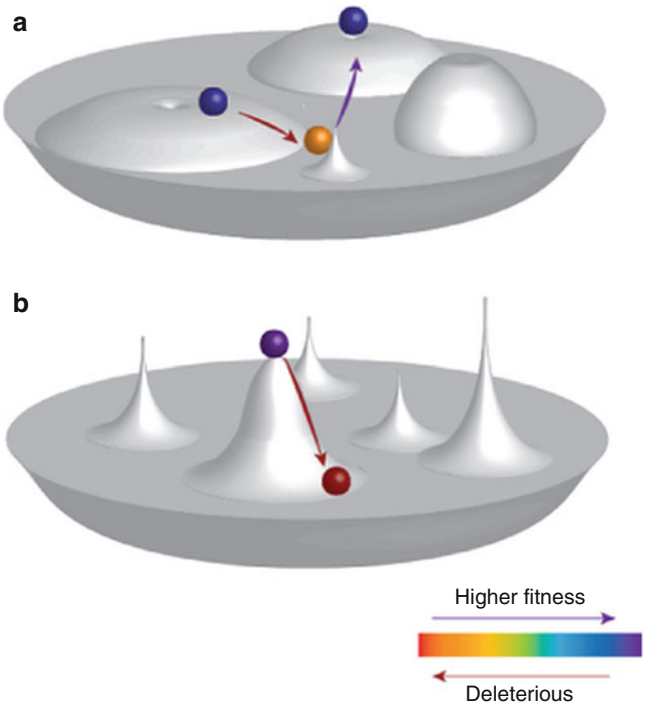

Fig. 7.2 Fitness landscape of an evolving population. Peaks represent regions of maximum relative biological fitness while valleys are regions of low fitness. In a smooth landscape (**a**) populations (*spheres*) can cross the valleys of low fitness without yielding lethal phenotypes (e.g., these populations go from high mean fitness to intermediate mean fitness). In rugged and complex landscape (**b**), crossing the valleys of low fitness is lethal and precludes populations from reaching new local fitness maxima through gradual evolution. This figure is taken from (Henderson et al. 2013), with author's permission

average fitness of the population increases gradually. When several adaptive mutations emerge, synergistic epistasis among them, that is interactions between mutations that increases the effects of single mutants on fitness in a non-linear fashion, leads to diminishing-returns epistasis: each mutation has lower beneficial effect for the individuals in the presence of another beneficial mutations than if it appeared alone in the ancestral genetic background (Chou et al. 2011; Khan et al. 2011; Kvitek and Sherlock 2011). Regardless of whether or not diminishing returns take place, beneficial mutations will lead populations to climb peaks in a fitness landscape (Fig. 7.2) (Orr 2009a, b). In the absence of interfering mutations, beneficial mutations will undergo refinement and selective sweep in the population (Atwood et al. 1951; Barrick and Lenski 2013). However, in asexual populations it is more frequent to observe cases in which the beneficial mutation needs to displace other beneficial mutations emerging during its fixation, thereby slowing down the fixation rate of

adaptive mutations. This effect, known as clonal interference (Fogle et al. 2008; Miralles et al. 1999), has been shown to be frequent in asexual populations of influenza (Strelkowa and Lassig 2012), the bacteriophage phiX174 (Pepin and Wichman 2008), bacteria (de Visser and Rozen 2006), and yeast (Kao and Sherlock 2008; Lang et al. 2013) but has only been characterized in yeast by deep sequencing yeast populations at frequent intervals (Lang et al. 2013).

Adaptive mutations need to be distinguished from those that are innovative, leading to new phenotypes adaptable to novel environments. Many research studies in this area have shown that such innovative mutations are often sudden and involve only one-to-few mutations. The identification of these mutations has been possible through the use of NGS, which has also enabled disentangling beneficial mutations from innovative ones. For example, in a recent study, Marchetti and colleagues showed that an experimentally evolved chimeric *Ralstonia solanacearum* strain, derived from a plant pathogen, could establish a symbiotic mutualistic association once evolved experimentally. This change in lifestyle occurred upon colonizing root nodules and was due to a single non-synonymous (amino acid replacing) mutation in the gene *hrpG* that encodes a protein regulating the expression of several virulent factors (Marchetti et al. 2010). In another study in which authors conducted a long-term evolution experiment with *E. coli* (LTEE), *E. coli* adapted to a glucose-limited medium, which also contained the bacterium-unusable citrate, evolved the ability to metabolize citrate after 30,000 generations in one of the 12 original replicate populations with which the experiment commenced (Blount et al. 2008). The emergence of this innovation required a single genome event in earlier generations (an enabling mutation), consistent on a chromosomal duplication that placed a transcription promoter upstream of a Citrate transporter-encoding gene (Blount et al. 2012).

The concept of genetic background and enabling mutations is very important to understand the term "evolvability"—the capacity of individuals or genotypes to evolve and adapt to a wide set of different conditions. Indeed, the combination of alleles existing in the population may well condition and constrain the evolutionary trajectories of new alleles, through either altering mutation rates or conditioning the nature and strength of epistatic interactions with new mutations (Meyer et al. 2012). The actual dynamics underlying the enabling effect of neutral mutation networks has been investigated in very simple systems, such as RNA folding (Wagner 2008), however, the role of enabling mutations versus compensatory mutations—those compensating the effects of destabilizing innovative mutations—remains the ground of intense investigation and debate.

As discussed earlier, populations with high mutation rates increase the per-capita chance of acquiring a beneficial mutation. In LTEE, the frequency of hyper-mutators is high, rising mutation rates 100-fold compared to that of the ancestral population (Mao et al. 1997). However, in recent studies it has been shown that hyper-mutators in experimental populations are generally followed by phenotypes with slow mutation rates, probably because such phenotypes prevent the loss of adaptive mutations in the populations and lower genetic load (Sniegowski et al. 2000; Wielgoss et al. 2013).

Convergent Evolution in Bacterial Experimental Populations

One of the most important questions yet unanswered is how repeatable is evolution. In particular, what is the role of contingency in the fixation of adaptive mutations? In a recent study (Tenaillon et al. 2012), authors evolved 115 *E. coli* populations for 2,000 generations of the bacterium to adapt to 42.2 °C, a complex environmental factor to which many pathways of the organism respond. To determine the diversity of adaptation of *E. coli* to high temperatures, they started the experiment from a single ancestral cell adapted to 37 °C. After 2,000 generations, the genome of one clone from each of the 115 experimentally evolving populations at 42.2 °C was sequenced. In addition to genome sequencing, the relative fitness of the evolved clones was assessed, observing a significant increase of fitness of the evolved strain at 42.2 °C in comparison with their ancestor. Interestingly, in 18 of the 115 lines, authors found a shared mutation in codon 966 of the RNA polymerase β-subunit (*rpoB*), and 17 lines contained an amino acid replacing mutation in codon 15 of the *rho* gene. In general, 20.2 % of genes mutated convergently in their experiment and 24.5 % of operons were convergently affected by mutations. This significant convergence was strongly driven by the epistatic interactions between new alleles. These experiments demonstrate that while the range of adaptive pathways may be bewildering, epistasis and genetic background can constrain the set of possible solutions to adapt to an environment, making evolution somewhat predictable.

Experimental Evolution Under Inefficient Natural Selection

To study the spectrum of mutations, researchers have evolved microbes, such as *E. coli* and *S. cerevisiae*, under controlled laboratory experiments and re-sequenced their genomes at different time points of the evolution experiment. Because the main objective of these experiments is to identify the breadth of mutations occurring in the genome, and calculate the rates of mutations, such populations have been evolved under very inefficient natural selection: replicates of evolving lines were single-colony transferred to new plates and this was repeated for hundreds or even thousands of generations (Fig. 7.3). These experiments have been useful to determine the spectrum and rate of mutations in *E. coli* (Lee et al. 2012) and *S. cerevisiae* (Lynch et al. 2008).

Purifying selection generally precludes the fixation of innovative mutations because they are generally destabilizing owing to the trade-off between current and novel adaptations (DePristo et al. 2005; Wilke et al. 2005; Zeldovich et al. 2007). There are a number of scenarios in which innovative mutations can be fixed under inefficient natural selection, including gene duplication (Ohno 1999), and systems with over-active mechanisms of mutational robustness, such as over-expressed molecular chaperones (Moran 1996).

How does gene duplication enable the fixation of innovative mutations? After the duplication of a gene, the two daughter copies are virtually identical, hence

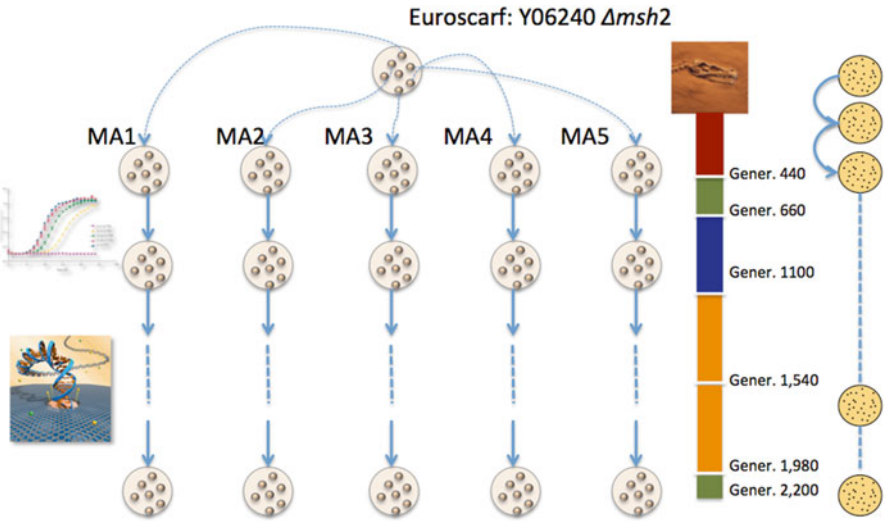

Fig. 7.3 Experimental evolution of populations of yeast under inefficient natural selection. Many replicate populations derive from a single yeast cell. To impose population bottlenecks and genetic drift a single colony is transferred to the new environment (plate). In the figure example, five independent lines of evolution started and evolved for many generations. At specific points of the evolution experiment, whole-genome sequencing and growth curves are conducted and mutations mapped in the reference genome

functionally redundant, with some exceptions that include non-duplicated regulatory elements, moving of one gene copy to a differently transcribed genome region or allele ancestral polymorphism (Lynch and Katju 2004). Such exceptions may well determine the spectrum of subsequent mutations of each gene copy, and consequently the functional fates of duplicates. The asymmetry between gene copies is avoided in many biological systems such as yeast through whole-genome duplication (WGD) but not through small-scale duplications (SSD). Accordingly, a number of studies have shown that the mechanism of duplication can determine the persistence of genes in duplicate, with WGDs being more prevalent among central genes in the network (although with some exceptions depending on the organism (Alvarez-Ponce and Fares 2012)), they are refractory to subsequent SSD events and dosage sensitive (Carretero-Paulet and Fares 2012; Conant and Wolfe 2006; Fares et al. 2013; Hakes et al. 2007; Makino and McLysaght 2010). These studies have shown that SSDs are more likely to present redundancy, hence mutational robustness and evolvability (Draghi et al. 2010), than WGDs. In particular, Fares and colleagues conducted a simple mutation accumulation experiment in which five lines of *S. cerevisiae* haploid strains derived from a single ancestor deficient in a mismatch repair gene (*msh2*) were evolved independently under strong genetic drift. They passaged these lines periodically by single colony transfers from one generation to the next for 2,200 generations. The whole genome of one colony was sequenced from each line and the distribution of non-synonymous SNPs in duplicates and singletons identified. As predicted by theory, SSDs showed significantly

larger number of non-synonymous SNPs than singletons and WGDs, supporting larger redundancy for SSDs than WGDs (Fares et al. 2013).

Experimental evolution has also been used to determine the role of a molecular chaperone in ameliorating the effects of deleterious non-lethal mutations. In an experiment in which several independent *E. coli* lines were subjected to single-colony passages, authors assessed the fitness of evolved population by competing them head-to-head to their ancestral population. After 3,200 generations of experimental bottlenecked evolution, cells presented half as much fitness as their ancestors owing to the increase in the deleterious mutational load owing to strong genetic drift effects. Over-expression of GroEL, a molecular chaperone essential in *E. coli* and which folds other proteins in the cell (Fayet et al. 1989; Lin and Rye 2006), allowed the recovery of about 88 % of the fitness of evolved cells (Fares et al. 2002). Interestingly, *groESL*, the operon encoding the chaperonin GroEL and its cofactor GroES, is abundantly synthesized in endosymbiotic mutualistic bacteria (Ahn et al. 1994; Aksoy 1995) that undergo strong genetic drift during their clonal transmission from mother host to the offspring (Buchner 1965). Experimental evolution of *E. coli* under inefficient natural selection reproduced therefore the transmission of endosymbiotic bacteria and identified GroEL as a mechanism of robustness against deleterious non-lethal mutations.

Concluding Remarks

Experimental evolution is a powerful tool to reproduce particular evolutionary processes with high repeatability and under tightly controlled environmental conditions. When combined with whole-genome sequencing, experimental evolution can inform on the dynamics underlying adaptations, speed of evolution, role of environment, and evolvability. Current studies have unveiled unprecedented and unexpected outcomes and have revealed complex dynamics to adaptation. While the general principles of evolution by natural selection clearly follow Darwinian laws, the evolutionary trajectories, contingency, constraints, and evolvability of organisms remain largely obscure. Future research in population genomics combined with NGS will be the key for understanding how do adaptations come about, how they interact, and where they lead.

References

Ahn TI, Lim ST, Leeu HK, Lee JE, Jeon KW (1994) A novel strong promoter of the groEx operon of symbiotic bacteria in *Amoeba proteus*. Gene 148:43–49

Aksoy S (1995) Molecular analysis of the endosymbionts of tsetse flies: 16S rDNA locus and over-expression of a chaperonin. Insect Mol Biol 4:23–29

Alvarez-Ponce D, Fares MA (2012) Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein–protein interaction network. Genome Biol Evol 4:1263–1274. doi:10.1093/gbe/evs101

Atwood KC, Schneider LK, Ryan FJ (1951) Periodic selection in *Escherichia coli*. Proc Natl Acad Sci U S A 37:146–155
Barrick JE, Lenski RE (2013) Genome dynamics during experimental evolution. Nat Rev Genet. doi:10.1038/nrg3564
Bennett AF, Lenski RE (2007) An experimental test of evolutionary trade-offs during temperature adaptation. Proc Natl Acad Sci U S A 104(Suppl 1):8649–8654. doi:10.1073/pnas.0702117104
Blount ZD, Borland CZ, Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. Proc Natl Acad Sci U S A 105:7899–7906. doi:10.1073/pnas.0803151105
Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental *Escherichia coli* population. Nature 489:513–518. doi:10.1038/nature11514
Bronikowski AM, Bennett AF, Lenski RE (2001) Evolutionary adaptation to temperature. VIII Effects of temperature on growth rate in natural isolates of Escherichia coli and Salmonella enterica from different thermal environments. Evolution 55:33–40
Buchner P (1965) Endosymbiosis of animals with plant microorganisms. Wiley, New York
Burke MK (2012) How does adaptation sweep through the genome? Insights from long-term selection experiments. Proc Biol Sci 279:5029–5038. doi:10.1098/rspb.2012.0799
Burke MK, Long AD (2012) What paths do advantageous alleles take during short-term evolutionary change? Mol Ecol 21:4913–4916
Burke MK, Dunham JP, Shahrestani P, Thornton KR, Rose MR, Long AD (2010) Genome-wide analysis of a long-term evolution experiment with *Drosophila*. Nature 467:587–590. doi:10.1038/nature09352
Carretero-Paulet L, Fares MA (2012) Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications. Mol Biol Evol 29:3541–3551. doi:10.1093/molbev/mss162
Chou HH, Chiu HC, Delaney NF, Segre D, Marx CJ (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. Science 332:1190–1192. doi:10.1126/science.1203799
Conant GC, Wolfe KH (2006) Functional partitioning of yeast co-expression networks after genome duplication. PLoS Biol 4:e109. doi:10.1371/journal.pbio.0040109
Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Myrray (Ed). London
de Visser JA, Rozen DE (2006) Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. Genetics 172:2093–2100. doi:10.1534/genetics.105.052373
DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet 6:678–687. doi:10.1038/nrg1672
Draghi JA, Parsons TL, Wagner GP, Plotkin JB (2010) Mutational robustness can facilitate adaptation. Nature 463:353–355. doi:10.1038/nature08694
Dykhuizen DE (1998) Santa Rosalia revisited: why are there so many species of bacteria? Antonie Van Leeuwenhoek 73:25–33
Elena SF, Lenski RE (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat Rev Genet 4:457–469. doi:10.1038/nrg1088
Fares MA, Ruiz-Gonzalez MX, Moya A, Elena SF, Barrio E (2002) Endosymbiotic bacteria: groEL buffers against deleterious mutations. Nature 417:398. doi:10.1038/417398a
Fares MA, Keane OM, Toft C, Carretero-Paulet L, Jones GW (2013) The roles of whole-genome and small-scale duplications in the functional specialization of *Saccharomyces cerevisiae* genes. PLoS Genet 9:e1003176. doi:10.1371/journal.pgen.1003176
Fayet O, Ziegelhoffer T, Georgopoulos C (1989) The groES and groEL heat shock gene products of *Escherichia coli* are essential for bacterial growth at all temperatures. J Bacteriol 171:1379–1385
Fogle CA, Nagle JL, Desai MM (2008) Clonal interference, multiple mutations and adaptation in large asexual populations. Genetics 180:2163–2173. doi:10.1534/genetics.108.090019
Futschik A, Schlotterer C (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207–218. doi:10.1534/genetics.110.114397
Futuyma DJ (1998) Evolutionary biology. Sinauer, Sunderland, MA

Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309:1387–1390. doi:10.1126/science.1112665

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappe MS, Short JM, Carrington JC, Mathur EJ (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309:1242–1245. doi:10.1126/science.1114057

Hakes L, Robertson DL, Oliver SG, Lovell SC (2007) Protein interactions from complexes: a structural perspective. Comp Funct Genomics 49356. doi: 10.1155/2007/49356

Hegreness M, Shoresh N, Hartl D, Kishony R (2006) An equivalence principle for the incorporation of favorable mutations in asexual populations. Science 311:1615–1617. doi:10.1126/science.1122469

Henderson B, Fares MA, Lund PA (2013) Chaperonin 60: a paradoxical, evolutionarily conserved protein family with multiple moonlighting functions. Biol Rev Camb Philos Soc 88:955–987. doi:10.1111/brv.12037

Kao KC, Sherlock G (2008) Molecular characterization of clonal interference during adaptive evolution in asexual populations of *Saccharomyces cerevisiae*. Nat Genet 40:1499–1504. doi:10.1038/ng.280

Kawecki TJ, Lenski RE, Ebert D, Hollis B, Olivieri I, Whitlock MC (2012) Experimental evolution. Trends Ecol Evol 27:547–560. doi:10.1016/j.tree.2012.06.001

Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF (2011) Negative epistasis between beneficial mutations in an evolving bacterial population. Science 332:1193–1196. doi:10.1126/science.1203801

Krogh A (1929) Progress of physiology. Am J Physiol 90:9

Kvitek DJ, Sherlock G (2011) Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. PLoS Genet 7:e1002056. doi:10.1371/journal.pgen.1002056

Lang GI, Rice DP, Hickman MJ, Sodergren E, Weinstock GM, Botstein D, Desai MM (2013) Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. Nature 500:571–574. doi:10.1038/nature12344

Lee MC, Chou HH, Marx CJ (2009) Asymmetric, bimodal trade-offs during adaptation of Methylobacterium to distinct growth substrates. Evolution 63:2816–2830. doi:10.1111/j.1558-5646.2009.00757.x

Lee H, Popodi E, Tang H, Foster PL (2012) Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci U S A 109:E2774–E2783. doi:10.1073/pnas.1210309109

Lenski RE, Rose MR, Simpson SC, Stadler SC (2000) Long-term experimental evolution in *Escherichia coli*. Am Nat 138:27

Lenski RE, Winkworth CL, Riley MA (2003) Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. J Mol Evol 56:498–508. doi:10.1007/s00239-002-2423-0

Lin Z, Rye HS (2006) GroEL-mediated protein folding: making the impossible, possible. Crit Rev Biochem Mol Biol 41:211–239. doi:10.1080/10409230600760382

Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20:544–549. doi:10.1016/j.tig.2004.09.001

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, Thomas WK (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. Proc Natl Acad Sci U S A 105:9272–9277. doi:10.1073/pnas.0803466105

Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci U S A 107:9270–9274. doi:10.1073/pnas.0914697107

Mao EF, Lane L, Lee J, Miller JH (1997) Proliferation of mutators in A cell population. J Bacteriol 179:417–422

Marchetti M, Capela D, Glew M, Cruveiller S, Chane-Woon-Ming B, Gris C, Timmers T, Poinsot V, Gilbert LB, Heeb P, Medigue C, Batut J, Masson-Boivin C (2010) Experimental evolution

of a plant pathogen into a legume symbiont. PLoS Biol 8:e1000280. doi:10.1371/journal.pbio.1000280
Meyer JR, Dobias DT, Weitz JS, Barrick JE, Quick RT, Lenski RE (2012) Repeatability and contingency in the evolution of a key innovation in phage lambda. Science 335:428–432. doi:10.1126/science.1214449
Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends Genet 17:589–596
Miralles R, Gerrish PJ, Moya A, Elena SF (1999) Clonal interference and the evolution of RNA viruses. Science 285:1745–1747
Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93:2873–2878
Ohno S (1999) Gene duplication and the uniqueness of vertebrate genomes circa 1970–1999. Semin Cell Dev Biol 10:517–522. doi:10.1006/scdb.1999.0332
Orr HA (2009a) Fitness and its role in evolutionary genetics. Nat Rev Genet 10:531–539. doi:10.1038/nrg2603
Orr HA (2009b) Testing natural selection. Sci Am 300:44–50
Ostrowski EA, Woods RJ, Lenski RE (2008) The genetic basis of parallel and divergent phenotypic responses in evolving populations of *Escherichia coli*. Proc Biol Sci 275:277–284. doi:10.1098/rspb.2007.1244
Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276:734–740
Pepin KM, Wichman HA (2008) Experimental evolution and genome sequencing reveal variation in levels of clonal interference in large populations of bacteriophage phiX174. BMC Evol Biol 8:85. doi:10.1186/1471-2148-8-85
Pikuta EV, Hoover RB, Tang J (2007) Microbial extremophiles at the limits of life. Crit Rev Microbiol 33:183–209. doi:10.1080/10408410701451948
Sniegowski PD, Gerrish PJ, Johnson T, Shaver A (2000) The evolution of mutation rates: separating causes from consequences. Bioessays 22:1057–1066. doi:10.1002/1521-1878(200012)22:12<1057::AID-BIES3>3.0.CO;2-W
Strelkowa N, Lassig M (2012) Clonal interference in the evolution of influenza. Genetics 192:671–682. doi:10.1534/genetics.112.143396
Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS (2012) The molecular diversity of adaptive convergence. Science 335:457–461. doi:10.1126/science.1212986
Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9:965–974. doi:10.1038/nrg2473
Wielgoss S, Barrick JE, Tenaillon O, Wiser MJ, Dittmar WJ, Cruveiller S, Chane-Woon-Ming B, Medigue C, Lenski RE, Schneider D (2013) Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. Proc Natl Acad Sci U S A 110:222–227. doi:10.1073/pnas.1219574110
Wilke CO, Bloom JD, Drummond DA, Raval A (2005) Predicting the tolerance of proteins to random amino acid substitution. Biophys J 89:3714–3720. doi:10.1529/biophysj.105.062125
Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol 3:e5. doi:10.1371/journal.pcbi.0030005

Chapter 8
Plant Carbohydrate Active Enzyme (CAZyme) Repertoires: A Comparative Study

Huansheng Cao, Alex Ekstrom, and Yanbin Yin

Why Do We Study Plant Cell Walls?

Lignocellulosic biofuels have drawn considerable attentions in the past decade for a number of reasons, such as (1) fossil-based fuel is not sustainable; (2) global warming is getting worse because of the increasing greenhouse gas emission from fossil-based fuel consumption and (3) starch-based biofuels compete with human foods. The study of the biosynthesis and modification of plant cell walls has thus become vitally important and timely because cell walls are the major components of biomass, the most promising renewable source for the production of biofuels and biomaterials (Ragauskas et al. 2006; Pauly and Keegstra 2008). In fact, other than being used as biofuel feedstock, plant cell walls also play other vital roles, such as providing structural support to cells and defense against pathogens, serving as cell-specific developmental and differentiation markers, and mediating or facilitating cell–cell communication, etc. Lastly, plant cell walls also have many economic uses in human and animal nutrition and as sources of natural textile fibers, paper and wood products, and components of fine chemicals and medicinal products. Clearly the study of plant cell walls is not only important and significant to the fundamental science but also has immense influence in our biotech industry and other economic fields.

H. Cao, Ph.D. • Y. Yin, Ph.D. (✉)
Department of Biological Sciences, Northern Illinois University,
325A Montgomery Hall, DeKalb, IL 60115-2857, USA
e-mail: yyin@niu.edu

A. Ekstrom
Department of Computer Science, Northern Illinois University, DeKalb, IL, USA

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_8

How CAZymes Are Related to Cell Wall Studies?

For decades the plant cell wall research community has been focusing on unraveling the complex chemical structures of cell walls using chemical and biochemical approaches. In addition, numerous studies have been published towards elucidating the metabolic and regulatory pathways for the syntheses of various cell wall biopolymers (celluloses (Lei et al. 2012; Endler and Persson 2011; Gu and Somerville 2010; Somerville 2006; Joshi and Mansfield 2007), hemicelluloses (Carpita 2012; Doering et al. 2012; York and O'Neill 2008; Scheller and Ulvskov 2010; Driouich et al. 2012; Sandhu et al. 2009), lignins (Carpita 2012; Zhong and Ye 2009; Vanholme et al. 2008; Boerjan et al. 2003; Weng and Chapple 2010; Xu et al. 2009; Humphreys and Chapple 2002; Li and Chapple 2010), and pectins (Driouich et al. 2012; Mohnen 2008; Harholt et al. 2010; Atmodjo et al. 2013)). A large number of genes have been experimentally characterized to be cell wall-related (CWR) and among them the most important are Carbohydrate Active enZymes (CAZymes) (Cantarel et al. 2009). For example, the Purdue Cell Wall genomics database collected ~1,000 Arabidopsis CWR genes (Yong et al. 2005) and the CAZy database (see below) annotated ~1,100 Arabidopsis genes to be CAZymes. A simple BLAST search suggests that 48 % of CWR genes are also CAZyme genes.

What Are CAZymes and the CAZyDB?

CAZymes are enzymes responsible for the synthesis, degradation, modification, and recognition of carbohydrates in all living organisms. According to the classification scheme established by the CAZy database (CAZyDB, www.cazy.org), CAZymes are classified into six different classes as of July 2013: GHs (glycoside hydrolases), GTs (glycosyltransferases), CEs (carbohydrate esterases), PLs (polysaccharide lyases), AAs (auxiliary activities), and CBMs (carbohydrate binding modules). During the past 20 years since early 1990s, the CAZyDB has created 132 GH families, 94 GT families, 16 CE families, 22 PL families, 10 AA families, and 67 CBM families. Among the six classes, GTs are used to build the glycosidic bonds to form polysaccharides and other glycol-molecules, GHs and PLs to break the glycosidic bonds and CEs to break ester bonds. CBMs are non-catalytic structural modules used for recognizing and binding different carbohydrates. AAs are most recently included in the CAZyDB to classify redox enzymes for degrading lignins, which were previously categorized by the FOLy database (Fungal Oxidative Lignin enzymes) (Levasseur et al. 2008).

Although CAZymes are found in all kinds of organisms, they are vitally important for plants because most of the energy and carbon fixed in the photosynthesis are turned into sugars and then to all kinds of carbohydrates by using CAZymes, which feed animals and microbes as foods and provide humans textiles, papers, fuels, industrial materials, etc. Since plants are the most carbohydrate-rich organisms, plant genomes typically encode thousands of CAZymes for carbohydrate metabolism.

Existing Studies on Plant CAZyomes

Two plant genomes have their CAZyome (all the CAZymes encoded in the genome) manually annotated in the CAZyDB: *Arabidopsis thaliana* has 1,167 CAZyme genes (http://www.cazy.org/e1.html) and *Oryza sativa* has 1,147 CAZyme genes (http://www.cazy.org/e4.html), accounting for 4.3 and 2.9 % of genes in Arabidopsis and Rice, respectively. These percentages were shown to be much higher than other eukaryotes such as yeast, fly and even higher than most bacteria (Coutinho et al. 2003; Henrissat et al. 2001). One of the reasons was because plants have many gene duplications in their genomes either at the whole-genome level or at local regions (Coutinho et al. 2003).

Although not included in the CAZyDB, the CAZyome of *Populus trichocarpa* has been published a few years ago (Geisler-Lee et al. 2006). It was shown that there are over 1,600 CAZymes in poplar, which is an underestimated number as CAZyme containing merely CBM domains were excluded in the count. Therefore, among the three studied plants poplar has the largest number of CAZymes, likely because poplar had a more recent whole genome duplication (chromosome polyploidy) compared with Arabidopsis and Rice (Geisler-Lee et al. 2006).

Other than the three model plants, the complete CAZyme gene repertories of other plants have not been surveyed. Nevertheless, as mentioned above, the Purdue Cell Wall genomics database has collected CWR gene families and compiled a CWR gene list for three plant species: Arabidopsis (Yong et al. 2005), Rice and Maize (Penning et al. 2009). Since the plant cell walls are mainly composed of various polysaccharides such as celluloses, hemicelluloses, and pectins, 48 % CWR genes encode CAZymes (see above). The non-CAZyme CWR genes include genes encoding lignin biosynthetic enzymes, sugar precursor synthases, transcription factors, transporters, signaling proteins, etc.

A paper published in 2004 (Yokoyama and Nishitani 2004) has compiled the CWR gene list for Arabidopsis and Rice, and reported 675 Arabidopsis genes of 32 protein families and 465 Rice genes of 20 protein families. The families with distinct sizes were compared between the two organisms using phylogenetic approaches and the size differences were attributed to the fact that dicot and monocot plants have cell walls with different carbohydrate compositions. For instance, Arabidopsis cell walls have more pectins and rice cell walls have xyloglucans with less complex side chains, in agreement with the finding that Arabidopsis appears to have more genes involved in pectin metabolism.

For early divergent non-seed plants, a recent study (Harholt et al. 2012) has looked at the GT gene repertories of spike moss *Selaginella moellendorffii* and moss *Physcomitrella patens* and used phylogenetic analyses to identify orthologs of different GT families. Interesting findings were made such as some GT families are absent in seed plants but are present in the spike moss and moss genomes. Cell wall chemical composition results were also referenced to help interpret the gene phylogenies in order to match the genotypes with phenotypes. For example, spike moss and moss both have GT2 proteins clustered with cellulose synthase (CesA) proteins

from cyanobacteria, red algae, and fungi, while the cluster does not contain any seed plant GT2 proteins. This suggests these lower land plants might have two sets of genes of distinct origins that are responsible for the synthesis of celluloses.

All these published comparative studies demonstrated that comparative genomics could be very useful to reveal new insights into the evolution and function of CAZymes and further be beneficial to a better understanding of the plant cell wall evolution and diversity.

dbCAN: A HMM Database for Large-Scale Analysis of CAZymes

To facilitate the research of CAZymes in any organisms including plants, we have built hidden Markov models (HMMs) for 330 CAZyme families created by the CAZyDB, which has enabled automated CAZyme annotation for fully sequenced genomes and metagenomes (Yin et al. 2012). The collection of HMMs is named dbCAN, a database for automated CAZyme annotation. Specifically, for each CAZyme family, we have an HMM to represent the multiple sequence alignment of the signature domains of CAZyDB-annotated proteins of that family.

Taking GT2 as an example, the signature domain of each CAZyme family is defined as a region that is conserved and present in all member proteins of the GT2 family. The GT2 member proteins are from GenBank, and were annotated by the CAZyDB to belong to the GT2 family. In order to locate the GT2 signature domain in all member proteins, we searched all the proteins against the NCBI conserved domain database (CDD) (Marchler-Bauer et al. 2009) to find a CDD model that are present in all (or most) of the query proteins, which in this case is pfam00535. We then extracted all the pfam00535 domains in the GT2 proteins and then built an HMM.

Note here the new GT2 HMM is different from the pfam00535 model as they are based on different training sets, with the former to be CAZyme GT2 family-specific as it is derived from GT2 proteins of the CAZyDB and the latter to be more general. Table 8.1 lists all of the CDD signature domains for CAZyme families. Note that there are 64 of them indicated as "self-built," meaning that no CDD model is found to cover most proteins in that CAZyme family. These families include 20 CBM families, 24 GH families, 11 GT families, 6 PL families, and 3 AA families. For half of the 64 families including all of the 20 CBM families, we were able to find the signature domain in at least one characterized member proteins by manually going through the literature that originally defined the family. We then scanned all the remaining member proteins to retrieve the signature domains and then built the HMMs. For the remaining families, since we do not know where the signature domains are, we had to use the CAZyDB-annotated full length proteins to build multiple sequence alignment and then manually remove long gaps and ambiguously aligned regions and then build HMMs. Fortunately all these families are catalytic enzyme families, which usually have long signature domains often covering most part of the full-length proteins.

Table 8.1 CAZyme families and their HMMs

CAZyme family	Origins of the signature domain HMMs in dbCAN	CAZyme family	Origins of the signature domain HMMs in dbCAN
AA0	NA	GH5	pfam00150
AA1	TIGR03389	GH50	pfam09206
AA10	pfam03067	GH51	COG3534
AA2	cd00692	GH52	pfam03512
AA3	COG2303	GH53	pfam07745
AA4	Self-built[a]	GH54	pfam09206
AA5	Self-built[a]	GH55	Self-built (pubmed: 17587693)
AA6	TIGR01755	GH56	pfam01630
AA7	COG0277	GH57	pfam03065
AA8	Self-built[a]	GH58	pfam12217
AA9	pfam03443	GH59	pfam02057
CBM0	NA	GH6	pfam01341
CBM1	pfam00734	GH60	Deleted family
CBM10	pfam02013	GH61	pfam03443
CBM11	pfam03425	GH62	pfam03664
CBM12	pfam02839	GH63	PRK10137
CBM13	pfam00652	GH64	Self-built[a]
CBM14	pfam01607	GH65	pfam03632
CBM15	pfam03426	GH66	Self-built[b]
CBM16	pfam02018	GH67	COG3661
CBM17	pfam03424	GH68	pfam02435
CBM18	smart00270	GH69	Replaced by PL16
CBM19	pfam03427	GH7	pfam00840
CBM2	pfam00553	GH70	pfam02324
CBM20	pfam00686	GH71	pfam03659
CBM21	pfam03370	GH72	pfam03198
CBM22	pfam02018	GH73	pfam01832
CBM23	pfam03425	GH74	smart00602
CBM24	Self-built (pubmed: 10636904)	GH75	pfam07335
CBM25	pfam03423	GH76	pfam03663
CBM26	Self-built (pubmed: 16230347)	GH77	pfam02446
CBM27	pfam09212	GH78	pfam05592
CBM28	pfam03424	GH79	pfam03662
CBM29	Self-built (pubmed: 11560933)	GH8	pfam01270
CBM3	pfam00942	GH80	cd00978
CBM30	pfam02927	GH81	pfam03639
CBM31	pfam11606	GH82	COG5434
CBM32	pfam00754	GH83	pfam00423

(continued)

Table 8.1 (continued)

CAZyme family	Origins of the signature domain HMMs in dbCAN	CAZyme family	Origins of the signature domain HMMs in dbCAN
CBM33	pfam03067	GH84	pfam07555
CBM34	cd02857	GH85	pfam03644
CBM35	pfam03422	GH86	Self-built[a]
CBM36	pfam03422	GH87	pfam00754
CBM37	smart00060	GH88	pfam07470
CBM38	Self-built (pubmed: 15109727)	GH89	pfam05089
CBM39	Self-built (pdb: 2RQE)	GH9	pfam00759
CBM4	pfam02018	GH90	pfam09251
CBM40	pfam02973	GH91	Self-built[a]
CBM41	pfam03714	GH92	pfam07971
CBM42	pfam05270	GH93	Self-built (pubmed: 14988022)
CBM43	pfam07983	GH94	COG3459
CBM44	pfam00801	GH95	Self-built (pubmed: 15262925)
CBM45	Self-built (pubmed: 16584202)	GH96	Self-built (pubmed: 17513582)
CBM46	pfam03442	GH97	pfam10566
CBM47	pfam00754	GH98	pfam08306
CBM48	pfam02922	GH99	Self-built[a]
CBM49	pfam09478	GT0	NA
CBM5	pfam02839	GT1	COG1819
CBM50	pfam01476	GT10	pfam00852
CBM51	pfam08305	GT11	pfam01531
CBM52	pfam10645	GT12	pfam00535
CBM53	pfam03423	GT13	pfam03071
CBM54	Self-built (pubmed: 19389758)	GT14	pfam02485
CBM55	Self-built (pubmed: 12011021)	GT15	pfam01793
CBM56	Self-built (pubmed: 9738929)	GT16	pfam05060
CBM57	pfam11721	GT17	pfam04724
CBM58	Self-built (pubmed: 20159465)	GT18	Self-built[b]
CBM59	Self-built (genbank: ACJ48973)	GT19	pfam02684
CBM6	pfam03422	GT2	pfam00535
CBM60	Self-built (pubmed: 20659893)	GT20	pfam00982
CBM61	Self-built (pubmed: 20826814)	GT21	cd02520
CBM62	Self-built (pubmed: 21454512)	GT22	pfam03901

(continued)

Table 8.1 (continued)

CAZyme family	Origins of the signature domain HMMs in dbCAN	CAZyme family	Origins of the signature domain HMMs in dbCAN
CBM63	Self-built (genbank: 110590882)	GT23	pfam05830
CBM64	Self-built (genbank: ADN02998.1)	GT24	cd06432
CBM65	Self-built (pdb: 4AEM)	GT25	pfam01755
CBM66	Self-built (pdb: 4AZZ)	GT26	pfam03808
CBM67	Self-built (genbank: BAC68538.1)	GT27	cd02510
CBM7	Deleted family	GT28	pfam04101
CBM8	Self-built (pubmed: 1447151)	GT29	pfam00777
CBM9	pfam06452	GT3	pfam05693
CE0	NA	GT30	pfam04413
CE1	pfam00756	GT31	pfam01762
CE10	COG0657	GT32	pfam04488
CE11	pfam03331	GT33	cd03816
CE12	cd01821	GT34	pfam05637
CE13	pfam03283	GT35	pfam00343
CE14	pfam02585	GT36	Replaced by GH94
CE15	pfam05448	GT37	pfam03254
CE16	cd01846	GT38	pfam07388
CE2	cd01831	GT39	pfam02366
CE3	cd01833	GT4	pfam00534
CE4	pfam01522	GT40	cd04186
CE5	pfam01083	GT41	COG3914
CE6	pfam03629	GT42	pfam06002
CE7	pfam05448	GT43	pfam03360
CE8	pfam01095	GT44	pfam04488
CE9	cd00854	GT45	pfam00535
GH0	NA	GT46	Deleted family
GH1	pfam00232	GT47	pfam03016
GH10	pfam00331	GT48	pfam02364
GH100	pfam04853	GT49	Self-built[a]
GH101	Self-built (pubmed: 16141207)	GT5	cd03791
GH102	pfam03562	GT50	pfam05007
GH103	TIGR02283	GT51	pfam00912
GH104	cd00736	GT52	pfam07922
GH105	pfam07470	GT53	pfam04602
GH106	Self-built[a]	GT54	pfam04666
GH107	Self-built (pubmed: 16880504)	GT55	pfam09488

(continued)

Table 8.1 (continued)

CAZyme family	Origins of the signature domain HMMs in dbCAN	CAZyme family	Origins of the signature domain HMMs in dbCAN
GH108	pfam05838	GT56	pfam07429
GH109	pfam01408	GT57	pfam03155
GH11	pfam00457	GT58	pfam05208
GH110	Self-built[b]	GT59	pfam04922
GH111	Self-built[a]	GT6	pfam03414
GH112	pfam09508	GT60	pfam11397
GH113	Self-built[a]	GT61	pfam04577
GH114	pfam03537	GT62	pfam03452
GH115	Self-built[a]	GT63	Self-built (pdb: 1BGT)
GH116	pfam04685	GT64	pfam09258
GH117	pfam04616	GT65	pfam10250
GH118	PRK11557	GT66	pfam02516
GH119	Self-built[a]	GT67	PTZ00210
GH12	pfam01670	GT68	pfam10250
GH120	pfam07602	GT69	pfam11735
GH121	Self-built[a]	GT7	pfam02709
GH122	COG4697	GT70	Self-built[a]
GH123	Self-built[a]	GT71	pfam11051
GH124	GH124	GT72	pfam11440
GH125	pfam06824	GT73	PRK09822
GH126	Self-built (genbank: ABG82272.1)	GT74	Self-built[a]
GH127	pfam07944	GT75	pfam03214
GH128	pfam11790	GT76	pfam04188
GH129	Self-built[a]	GT77	pfam03407
GH13	pfam00128	GT78	Self-built[a]
GH130	pfam04041	GT79	Self-built[a]
GH131	Self-built (genbank: AFQ89876.1)	GT8	pfam01501
GH132	Self-built (genbank: EAL86926.1)	GT80	pfam11477
GH14	pfam01373	GT81	PRK13915
GH15	pfam00723	GT82	pfam06306
GH16	cd00413	GT83	COG1807
GH17	pfam00332	GT84	pfam10091
GH18	pfam00704	GT85	pfam12250
GH19	cd00325	GT86	Deleted family
GH2	COG3250	GT87	pfam09594
GH20	pfam00728	GT88	Self-built[b]
GH21	Deleted family	GT89	Self-built[a]
GH22	pfam00062	GT9	pfam01075

(continued)

Table 8.1 (continued)

CAZyme family	Origins of the signature domain HMMs in dbCAN	CAZyme family	Origins of the signature domain HMMs in dbCAN
GH23	cd00254	GT90	smart00672
GH24	cd00737	GT91	pfam12141
GH25	pfam01183	GT92	pfam01697
GH26	pfam02156	GT93	Self-built (genbank: ZP_03542636.1)
GH27	pfam02065	GT94	Self-built (genbank: AAA86377)
GH28	pfam00295	PL0	NA
GH29	pfam01120	PL1	smart00656
GH3	pfam00933	PL10	pfam09492
GH30	COG5520	PL11	Self-built[b]
GH31	pfam01055	PL12	pfam07940
GH32	pfam00251	PL13	Self-built[a]
GH33	cd00260	PL14	Self-built[a]
GH34	pfam00064	PL15	pfam07940
GH35	pfam01301	PL16	pfam07212
GH36	COG3345	PL17	pfam07940
GH37	pfam01204	PL18	pfam08787
GH38	pfam01074	PL19	Replaced by GH91
GH39	pfam01229	PL2	pfam06917
GH4	pfam02056	PL20	Self-built[a]
GH40	Deleted family	PL21	pfam07940
GH41	Deleted family	PL22	TIGR02800
GH42	pfam02449	PL3	pfam03211
GH43	pfam04616	PL4	pfam09284
GH44	Self-built[b]	PL5	pfam05426
GH45	pfam02015	PL6	Self-built[a]
GH46	cd00978	PL7	pfam08787
GH47	pfam01532	PL8	pfam02278
GH48	pfam02011	PL9	Self-built[b]
GH49	pfam03718	–	–

[a]All member proteins in the family are used
[b]Only characterized member proteins are used

Such HMM representing protein families/domains has been widely used in other well-known protein family databases such as Pfam (Punta et al. 2012), TIGRFAMs (Haft et al. 2003), PANTHER (Mi et al. 2005), SUPERFAMILY (Wilson et al. 2009), etc. and proved to be very efficient in large-scale analyses such as genome annotation.

With regard to the search of HMMs, unlike another commonly used approach, the BLAST search that uses protein sequences as the query, the HMM-based search uses the HMM search tool HMMER 3.0 (http://hmmer.janelia.org/) (Eddy 2011), which takes the HMMs as the query and a protein sequence dataset (e.g., the entire plant proteome) as the database or vice versa. Since an HMM is a representation of

a multiple sequence alignment of a certain protein domain, it encompasses the common characteristics of all member sequences in the alignment and thus the HMM search is more sensitive compared to the BLAST search. The result will report the occurrences of all the matched HMM domains and the boundaries of the domains, so that it is particularly well suited for annotating proteins having multiple domains (including repetitive domains).

With the 330 CAZyme HMMs, we have also built a web server (http://csbl.bmb.uga.edu/dbCAN/annotate.php) to allow users to submit their protein sequences of a whole genome for the automated and large-scale annotation of CAZymes (Yin et al. 2012).

CAZyome of Fully Sequenced Plants

Using the HMMER 3.0 as the tool and the 330 CAZyme HMMs as the query, we have scanned 31 fully sequenced plants and 2 chlorophytic green algae available in the Phytozome database (Goodstein et al. 2012). AA families were not included in this analysis as they became available only recently. The 31 plants include 23 dicots, 6 monocots, 1 spike moss, and 1 moss. We used *E*-value <1e-5 and HMM coverage >0.3 to keep the hits. The HMM coverage means the fraction of CAZyme HMM (representing the CAZyme domain) that is covered in the alignment with the hit protein. Our previous evaluation using annotated Arabidopsis CAZymes in the CAZyDB as the benchmark suggests that the coverage >0.3 is the best cutoff to balance the rates of false positives and false negatives and overall our automated CAZyme annotation achieves sensitivity = 96.3 % and PPV = 78.8 % for Arabidopsis (Yin et al. 2012).

As the result, Fig. 8.1 shows the total number of CAZymes found in the 33 surveyed genomes and the breakdown of the five different CAZyme classes. In agreement with previous papers (Coutinho et al. 2003), there is a positive correlation between the number of CAZymes (4th column) and the total number of proteins (3rd column) encoded in the genomes (Pearson's correlation coefficient $R = 0.77$, P-value = 1.8e-07), meaning that the more proteins encoded in the genome, the more CAZymes the genome has. Figure 8.1 also shows that this correlation largely remains for all of the five CAZyme classes, although the correlation coefficient R varies slightly with PL having the lowest R and GT has the highest R (GT: 0.78, GH: 0.74, CE: 0.72, CBM: 0.69 and PL: 0.58).

Since the correlation is not perfect, we further calculated the percentage of CAZymes in the 33 genomes by dividing the number CAZymes by the total number of proteins (including splicing isoforms). As shown in Fig. 8.2, the CAZyme % ranges from 2.1 % in *VolVox carteri* to 5.1 % in *Mimulus guttatus*. On average, there are 3.8 % CAZymes encoded in the plant genomes. Comparatively speaking, chlorophytic green algae have lower CAZyme %, possibly because they are only distantly related to land plants and might not have as much carbohydrate contents as plants. Moss also has a lower percentage (2.6 %), but spike moss has a higher CAZyme % (4.4 %). Most seed plants have higher CAZyme %, but there are some

Clade	Species	Total #	CAZyme #	CBM #	CE #	GH #	GT #	PL #
Dicot	*Manihot_Esculenta*	34151	1667	212	299	515	603	38
	Ricinus_Communis	31221	1258	153	251	391	434	29
	Linum_Usitatissimum	43484	2173	260	371	685	801	56
	Populus_Trichocarpa	45033	1923	297	353	537	695	41
	Medicago_Truncatula	53423	1223	115	254	367	454	33
	Phaseolus_Vulgaris	30721	1426	169	269	421	534	33
	Glycine_Max	73320	2491	268	478	751	942	52
	Cucumis_Sativus	30364	1091	119	207	348	393	24
	Prunus_Persica	28702	1363	154	259	418	504	28
	Malus_Domestica	63517	2362	277	431	667	925	62
	Arabidopsis_Thaliana	35386	1338	160	269	393	483	33
	Arabidopsis_Lyrata	32670	1337	160	258	407	478	34
	Capsella_Rubella	28447	1371	173	270	391	501	36
	Brassica_Rapa	41019	1949	226	389	579	695	60
	Thellungiella_Halophila	29284	1362	171	242	423	513	13
	Carica_Papaya	27793	902	103	168	301	310	20
	Gossypium_Raimondii	77267	1755	231	305	508	642	69
	Citrus_Sinensis	46147	1969	204	335	626	758	46
	Citrus_Clementina	35976	1635	166	307	497	630	35
	Eucalyptus_Grandis	54935	1799	219	293	511	750	26
	Vitis_Vinifera	26346	1181	143	225	390	400	23
	Mimulus_Guttatus	28282	1453	183	294	429	514	33
	Aquilegia_Coerulea	41063	1204	118	246	367	448	25
Monocot	*Sorghum_Bicolor*	29448	1425	145	284	394	587	15
	Zea_Mays	63540	2259	244	367	684	934	30
	Setaria_Italica	40599	1750	185	344	496	709	16
	Panicum_Virgatum	70071	2889	264	553	817	1226	29
	Oryza_Sativa	66338	1473	169	278	421	590	15
	Brachypodium_Distachyon	31029	1326	137	259	368	551	11
Spike moss	*Selaginella_Moellendorffii*	22285	980	126	203	266	370	15
Moss	*Physcomitrella_Patens*	38354	1014	114	159	307	406	28
Chlorophytic green algae	*Chlamydomonas_Reinhardtii*	17114	483	91	42	86	261	3
	Volvox_Carteri	15285	326	79	29	54	162	2
	Total	1332614	50157	5835	9291	14815	19203	1013

Fig. 8.1 The 33 plant and green algae and the number of CAZymes

Clade	Species	CBM% * 100	CE% * 100	GH% * 100	GT% * 100	PL% * 100	CAZyme%
Dicot	*Manihot_Esculenta*	0.62077245	0.87552341	1.50800855	1.76568768	0.11127053	0.04881263
	Ricinus_Communis	0.49005477	0.80394606	1.25236219	1.39009	0.0928862	0.04029339
	Linum_Usitatissimum	0.59792107	0.85318738	1.57529206	1.84205685	0.128783	0.0499724
	Populus_Trichocarpa	0.65951635	0.78386961	1.19245886	1.54331268	0.09104435	0.04270202
	Medicago_Truncatula	0.21526309	0.47545065	0.68697003	0.84982124	0.06177115	0.02289276
	Phaseolus_Vulgaris	0.5501123	0.87562254	1.3703981	1.73822467	0.10741838	0.04641776
	Glycine_Max	0.365521	0.65193672	1.02427714	1.28477905	0.07092199	0.03397436
	Cucumis_Sativus	0.39191147	0.68172836	1.14609406	1.29429588	0.07904097	0.03593071
	Prunus_Persica	0.53654798	0.90237614	1.45634451	1.75597519	0.09755418	0.04748798
	Malus_Domestica	0.43610372	0.6785585	1.05011257	1.45630304	0.09761166	0.03718689
	Arabidopsis_Thaliana	0.45215622	0.76018764	1.11060872	1.36494659	0.09325722	0.03781156
	Arabidopsis_Lyrata	0.48974594	0.78971534	1.24579125	1.46311601	0.10407101	0.0409244
	Capsella_Rubella	0.60814849	0.94913348	1.37448589	1.76116989	0.12655113	0.04819489
	Brassica_Rapa	0.55096419	0.94834101	1.41154099	1.69433677	0.14627368	0.04751457
	Thellungiella_Halophila	0.58393662	0.82638984	1.4444748	1.75180986	0.04439284	0.04651004
	Carica_Papaya	0.37059691	0.60446875	1.08300651	1.11538877	0.07196057	0.03245422
	Gossypium_Raimondii	0.29896333	0.39473514	0.65746049	0.83088511	0.08930074	0.02271345
	Citrus_Sinensis	0.44206557	0.72594101	1.35653455	1.64257698	0.09968145	0.042668
	Citrus_Clementina	0.46141872	0.85334668	1.38147654	1.75116744	0.09728708	0.04544696
	Eucalyptus_Grandis	0.39865295	0.5333576	0.93019022	1.36524984	0.04732866	0.03274779
	Vitis_Vinifera	0.54277689	0.85401959	1.48030061	1.51825704	0.08729978	0.04482654
	Mimulus_Guttatus	0.64705466	1.03953044	1.51686585	1.81741037	0.11668199	0.05137543
	Aquilegia_Coerulea	0.28736332	0.59907946	0.89374863	1.0910065	0.06088206	0.0293208
Monocot	*Sorghum_Bicolor*	0.49239337	0.96441184	1.33795164	1.9933442	0.05093725	0.04839038
	Zea_Mays	0.38401007	0.57758892	1.07648725	1.4699402	0.04721435	0.03555241
	Setaria_Italica	0.45567625	0.84731151	1.22170497	1.74634843	0.03940984	0.04310451
	Panicum_Virgatum	0.37676071	0.78919953	1.16596024	1.74965392	0.04138659	0.04122961
	Oryza_Sativa	0.25475595	0.419066	0.63462872	0.88938467	0.02261147	0.02220447
	Brachypodium_Distachyon	0.44152245	0.83470302	1.1859873	1.77575816	0.03545071	0.04273422
Spike moss	*Selaginella_Moellendorffii*	0.56540274	0.91092663	1.193628	1.66030963	0.06730985	0.04397577
Moss	*Physcomitrella_Patens*	0.29723106	0.41455911	0.80043802	1.05855973	0.07300412	0.02643792
Chlorophytic green algae	*Chlamydomonas_Reinhardtii*	0.53172841	0.24541311	0.50251256	1.5250672	0.01752951	0.02822251
	Volvox_Carteri	0.51684658	0.18972849	0.35328754	1.05986261	0.01308472	0.0213281
	Total	0.43786123	0.69720114	1.11172478	1.44100242	0.07601601	0.03763806

Fig. 8.2 The percentage of CAZymes

exceptions such as *Medicago_Truncatula* (2.3 %), *Gossypium_Raimondii* (2.3 %), *Aquilegia_Coerulea* (2.9 %), and *Oryza_Sativa* (2.2 %).

If we look at the different CAZyme classes, more interesting findings emerge. For example, the green algae have very low PL %, CE %, and GH %, but fairly high CBM % compared with land plants. Monocots seem to have lower PL % and CBM % than dicots.

We also looked into each CAZyme family to compare the numbers across different plants. Figures 8.3, 8.4, 8.5, 8.6, and 8.7 show the distribution of different CAZymes in different genomes: GH families (Fig. 8.3), GT families (Fig. 8.4), CE families (Fig. 8.5), PL families (Fig. 8.6), and CBM families (Fig. 8.7). These heat maps provide very general overviews of the presence/absence and the abundance of each CAZyme families. In-depth comparative analyses such as phylogenetic analysis and expression analysis by referring to existing knowledge about the gene functions in the literature will be needed to make any substantive conclusions. Here we omit the details and only summarize some of the most obvious observations by visually inspecting these plots.

GHs There are 64 GH families having member proteins in at least one plant or green algal genomes. The top families include GH28, GH17, GH1, GH16, which are all reported to be CWR (Yokoyama and Nishitani 2004). However, about half of the 64 families have only very few members (e.g., <5). Some of these families are only found in green algae, e.g. GH125, GH114, GH99, and others are only sporadically distributed in a small number of plants, such as GH117, GH113, GH106, GH105, GH104, GH103, GH97, GH92, GH78, GH76, GH71, GH55, GH54, GH53, GH36, GH24. Further examination of the HMMER results is needed to determine whether these families are indeed present in these genomes or not. If they are, then it remains to be answered why they are only found in very few plants, where are they originated from, were they transferred from other organisms like bacteria or were they lost in other plants in the evolution?

GT 60 GT families have genes in at least one of the surveyed genomes. The top families with higher number of member genes include GT47, GT8, GT1, GT2, GT4, GT31, GT77, GT14, all of which except GT1 and GT14 are involved in cell wall synthesis (Yokoyama and Nishitani 2004). Comparatively, it appears that GT34, GT37, GT43, GT61, GT77 families have more genes in monocots than in dicots, while GT75 and GT8 have more genes in dicots than in monocots, and GT47 and GT77 have higher percentages of genes in green algae than in land plants.

Like GH families, there are also many families present in a small number of genomes. For example, GT94, GT71, GT69, GT60, GT49, GT25, GT15 are only found in green algae. GT81 and GT78 are only found in moss and spike moss. GT51 was also found to be only present in moss and spike moss (Harholt et al. 2012), but here we show that they appear to be present in nine seed plants as well. There are also families only absent in green algae or in moss or spike moss, such as GT58, GT51, GT43, GT29, GT22, GT21, GT9, etc. But again, further sequence analyses are needed to confirm this and to answer why and how this happens.

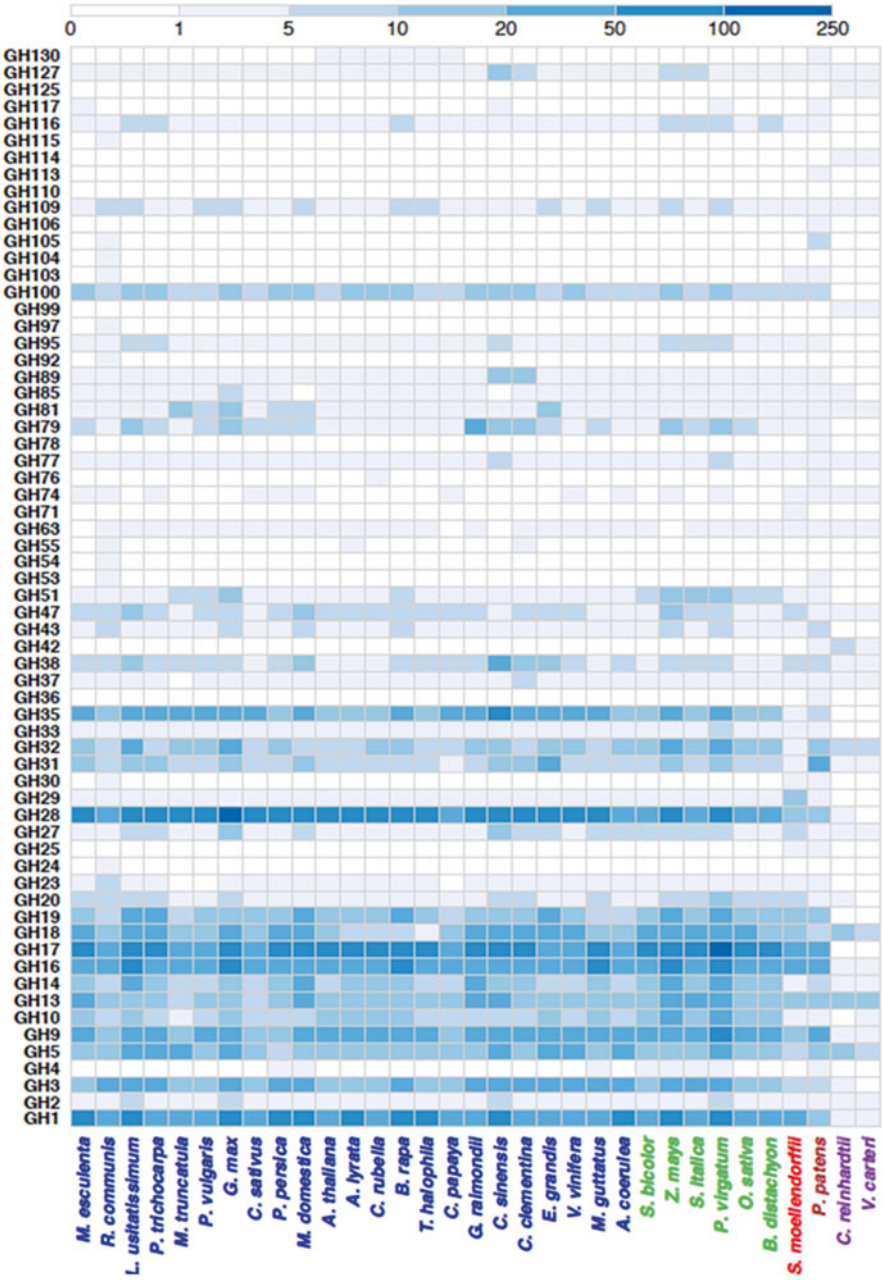

Fig. 8.3 Heat map of GH family sizes

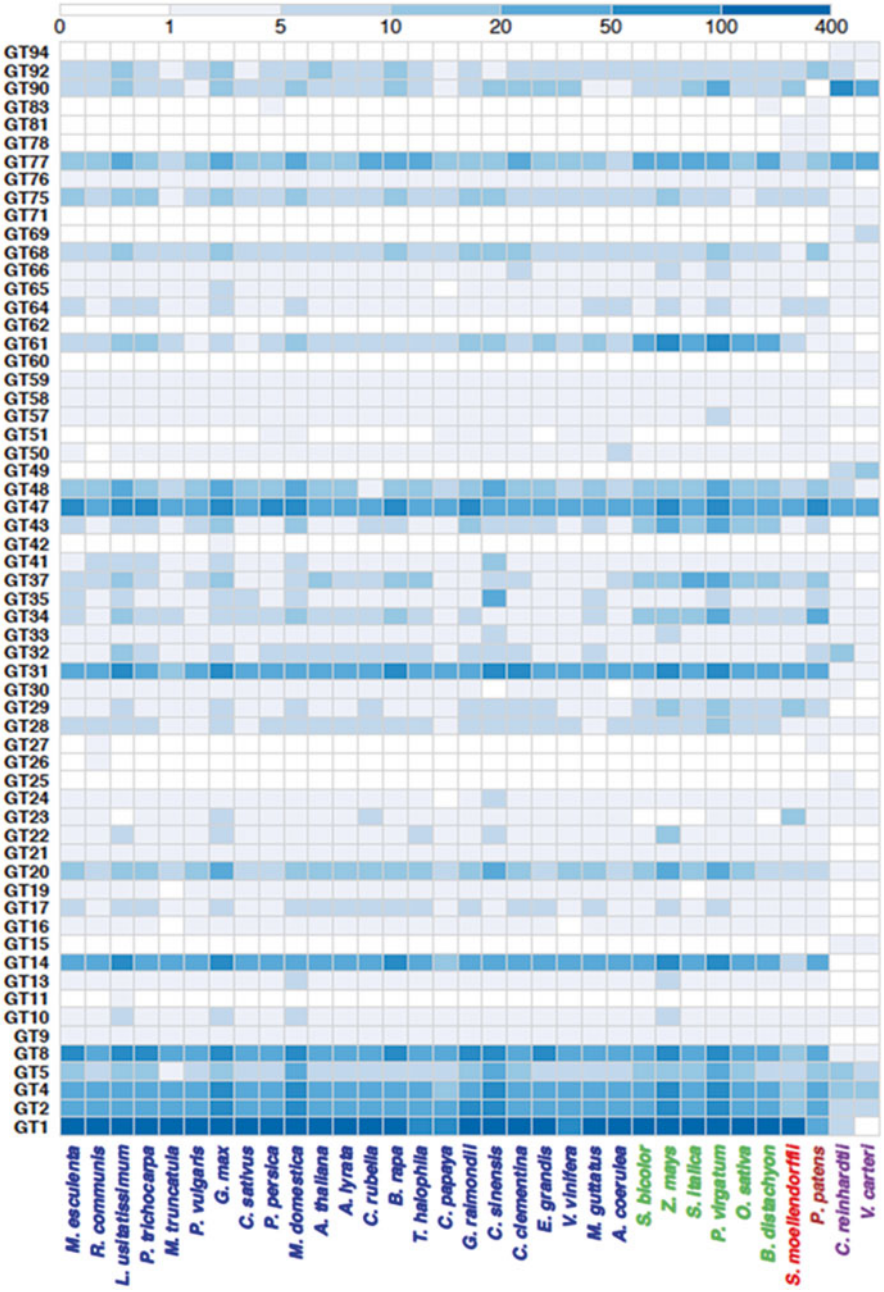

Fig. 8.4 Heat map of GT family sizes

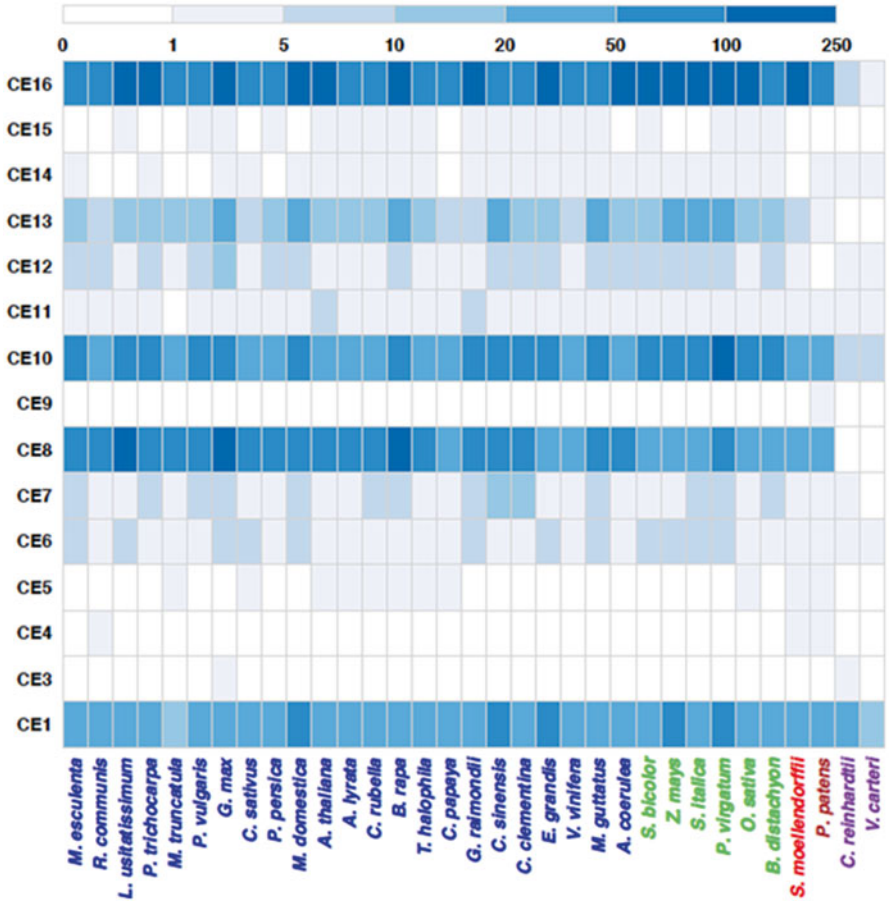

Fig. 8.5 Heat map of CE family sizes

CE Except for CE2, all of the 16 CE families have member genes in plants and algae. The top families are CE16, CE10, CE8, and CE1. In contrast, CE9 is only found in moss, CE3 and CE4 are only found in two and three species, respectively. CE8 and CE13 genes encode pectin methylesterases and pectin acetylesterases and they both are absent in green algae. CE8 seems to have higher number of genes in dicots than monocots.

PL The two pectin lyases families, PL1 and PL4 are widely found in land plants but are absent in green algae. Both families appear to have higher percentages of genes in dicots compared to monocots. Other PL families are only sporadically distributed in a small number of species and have a very small number of genes.

CBM Almost half of the 67 CBM families are found in plants and algae. The top families are CBM57, CBM43, CBM50, CBM48, CBM22, CBM20, CBM18.

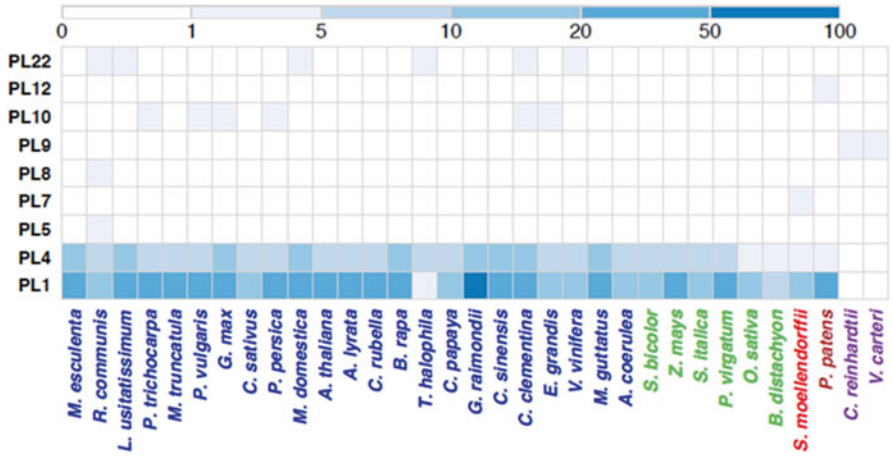

Fig. 8.6 Heat map of PL family sizes

The following families are widely distributed in land plants but are missing in chlorophytic green algae: CBM57, CBM43, CBM49, CBM22, CBM16, as opposed to CBM47, CBM21, CBM14, which have genes only in the algae.

Phylogenetic Analysis Is Useful to the Study of the Function and Evolution of CAZymes

Again further detailed phylogenetic analyses will be very useful to delineate the function and evolution of these CAZyme families, like what we did previously on cellulose and hemicellulose backbone synthases of GT2 (Yin et al. 2009), xylan and pectin synthesis-related GT8, 43, 47 (Kong et al. 2009, 2011; Kulkarni et al. 2011; Yin et al. 2010, 2011a), NDP-sugar inter-conversion enzymes (Yin et al. 2011b; Gu et al. 2010), monolignol synthetic enzymes (Escamilla-Trevino et al. 2010; Zhao et al. 2010), and transcription factors (Shen et al. 2009). As an example, Fig. 8.8 shows a phylogeny of GT2 proteins from 16 plant and two chlorophytic green algal genomes. It clearly presents the sequence-based clustering of GT2 proteins into many different clusters. By locating experimentally studied genes in the phylogeny and referring to published literature on the functional roles of these genes, we were able to further classify plant GT2 proteins into CesA subfamily and 10 cellulose synthase-like (Csl) subfamilies. The un-labeled branches in the phylogeny are less studied and are apparently non-Csl GT2 proteins. Comparing to previous studies (Yin et al. 2009), we were able to show that CslG and CslJ are not restricted to dicots and monocots, respectively, by surveying more completed genomes. Similar analyses will be conducted on other CAZyme families in the future.

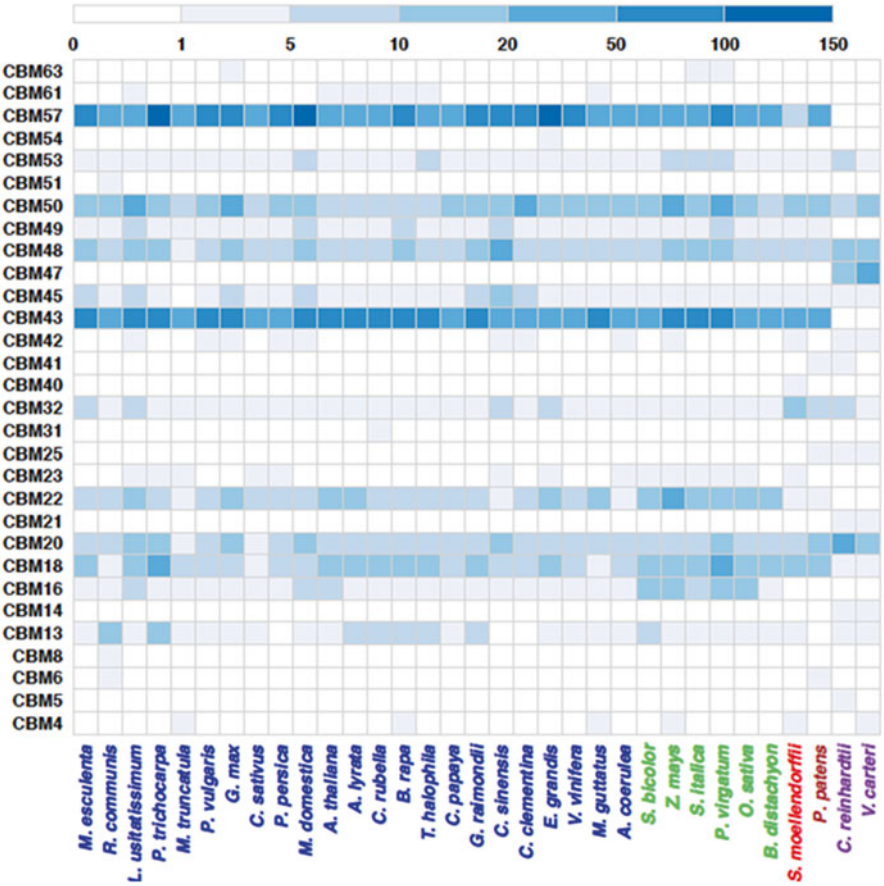

Fig. 8.7 Heat map of CBM family sizes

Future Development

A web-based database named PlantCAZyme is being developed in our lab (http://cys.bios.niu.edu/plantcazyme/index.php) to provide public access to the pre-computed plant CAZyme sequence and annotation data. Based on CAZyme sequences, secondary data will be derived through more in-depth bioinformatics analyses, such as computer-based functional annotation, sequence alignment, phylogenetic trees, predicted protein structures, orthologs and paralogs, genomic locations, etc. A web-based BLAST search function will also be available to allow user to search against the pre-computed plant CAZyme sequence database.

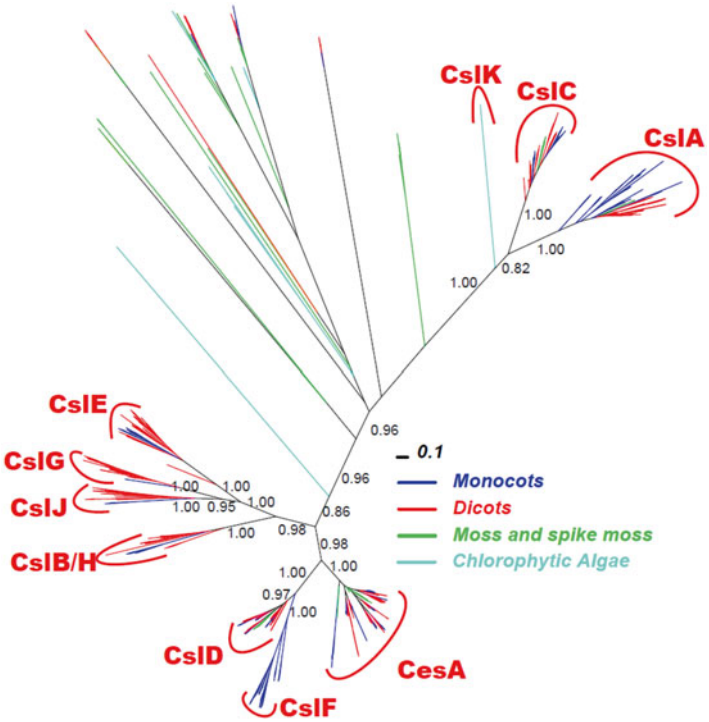

Fig. 8.8 Phylogeny of 849 GT2 proteins from 16 land plants and two green algae. The full-length protein sequences are used to build the phylogeny. The FastTree program (Price et al. 2010) is used to build this tree, with bootstrap values larger than 0.70 (70 %) are shown beside selected nodes forming the major Csl clusters. Csl clusters are labeled according to the presence of known Csl proteins in each cluster

References

Atmodjo MA, Hao Z, Mohnen D (2013) Evolving views of pectin biosynthesis. Annu Rev Plant Biol 64:747–779

Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Plant Biol 54:519–546

Cantarel BL et al (2009) The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37:D233–D238

Carpita NC (2012) Progress in the biological synthesis of the plant cell wall: new ideas for improving biomass for bioenergy. Curr Opin Biotechnol 23(3):330–337

Coutinho PM et al (2003) Why are there so many carbohydrate-active enzyme-related genes in plants? Trends Plant Sci 8(12):563–565

Doering A, Lathe R, Persson S (2012) An update on xylan synthesis. Mol Plant 5(4):769–771

Driouich A et al (2012) Golgi-mediated synthesis and secretion of matrix polysaccharides of the primary cell wall of higher plants. Front Plant Sci 3:79

Eddy SR (2011) Accelerated profile HMM searches. PLoS Comput Biol 7(10):e1002195

Endler A, Persson S (2011) Cellulose synthases and synthesis in Arabidopsis. Mol Plant 4(2):199–211

Escamilla-Trevino LL et al (2010) Switchgrass (Panicum virgatum) possesses a divergent family of cinnamoyl CoA reductases with distinct biochemical properties. New Phytol 185(1):143–155
Geisler-Lee J et al (2006) Poplar carbohydrate-active enzymes. Gene identification and expression analyses. Plant Physiol 140(3):946–962
Goodstein DM et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40(Database issue):D1178–D1186
Gu Y, Somerville C (2010) Cellulose synthase interacting protein: a new factor in cellulose synthesis. Plant Signal Behav 5(12):1571–1574
Gu XG et al (2010) Identification of a bifunctional UDP-4-keto-pentose/UDP-xylose synthase in the plant pathogenic bacterium Ralstonia solanacearum strain GMI1000, a distinct member of the 4,6-dehydratase and decarboxylase family. J Biol Chem 285(12):9030–9040
Haft DH, Selengut JD, White O (2003) The TIGRFAMs Database of Protein Families. Nucleic Acids Res 31(1):371–373
Harholt J, Suttangkakul A, Vibe Scheller H (2010) Biosynthesis of pectin. Plant Physiol 153(2):384–395
Harholt J et al (2012) The glycosyltransferase repertoire of the spikemoss *Selaginella moellendorffii* and a comparative study of its cell wall. PLoS One 7(5):e35846
Henrissat B, Coutinho PM, Davies GJ (2001) A census of carbohydrate-active enzymes in the genome of *Arabidopsis thaliana*. Plant Mol Biol 47(1–2):55–72
Humphreys JM, Chapple C (2002) Rewriting the lignin roadmap. Curr Opin Plant Biol 5(3):224–229
Joshi CP, Mansfield SD (2007) The cellulose paradox–simple molecule, complex biosynthesis. Curr Opin Plant Biol 10(3):220–226
Kong Y et al (2009) Two poplar glycosyltransferase genes, PdGATL1.1 and PdGATL1.2, are functional orthologs to PARVUS/AtGATL1 in Arabidopsis. Mol Plant 2(5):1040–1050
Kong Y et al (2011) Molecular analysis of a family of Arabidopsis genes related to galacturonosyltransferases. Plant Physiol 155(4):1791–1805
Kulkarni AR et al (2011) The ability of land plants to synthesize glucuronoxylans predates the evolution of tracheophytes. Glycobiology 22(3):439–451
Lei L, Li S, Gu Y (2012) Cellulose synthase complexes: composition and regulation. Front Plant Sci 3:75
Levasseur A et al (2008) FOLy: an integrated database for the classification and functional annotation of fungal oxidoreductases potentially involved in the degradation of lignin and related aromatic compounds. Fungal Genet Biol 45(5):638–645
Li X, Chapple C (2010) Understanding lignification: challenges beyond monolignol biosynthesis. Plant Physiol 154(2):449–452
Marchler-Bauer A et al (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res 37(Database issue):D205–D210
Mi H et al (2005) The PANTHER Database of Protein Families, Subfamilies, Functions and Pathways. Nucleic Acids Res 33(Database issue):D284–D288
Mohnen D (2008) Pectin structure and biosynthesis. Curr Opin Plant Biol 11(3):266–277
Pauly M, Keegstra K (2008) Cell-wall carbohydrates and their modification as a resource for biofuels. Plant J 54(4):559–568
Penning BW et al (2009) Genetic resources for maize cell wall biology. Plant Physiol 151(4):1703–1728
Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490
Punta M et al (2012) The Pfam Protein Families Database. Nucleic Acids Res 40(Database issue):D290–D301
Ragauskas AJ et al (2006) The path forward for biofuels and biomaterials. Science 311(5760):484–489
Sandhu AP, Randhawa GS, Dhugga KS (2009) Plant cell wall matrix polysaccharide biosynthesis. Mol Plant 2(5):840–850

Scheller HV, Ulvskov P (2010) Hemicelluloses. Annu Rev Plant Biol 61:263–289
Shen H et al (2009) A bioinformatic analysis of NAC genes for plant cell wall development in relation to lignocellulosic bioenergy production. Bioenergy Res 2(4):217–232
Somerville C (2006) Cellulose synthesis in higher plants. Annu Rev Cell Dev Biol 22:53–78
Vanholme R et al (2008) Lignin engineering. Curr Opin Plant Biol 11(3):278–285
Weng JK, Chapple C (2010) The origin and evolution of lignin biosynthesis. New Phytol 187(2):273–285
Wilson D et al (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37(Database issue):D380–D386
Xu Z et al (2009) Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom. BMC Bioinformatics, 10 Suppl 11, S3
Yin Y, Huang J, Xu Y (2009) The cellulose synthase superfamily in fully sequenced plants and algae. BMC Plant Biol 9(1):99
Yin Y et al (2010) Evolution and function of the plant cell wall synthesis-related glycosyltransferase family 8. Plant Physiol 153(4):1729–1746
Yin Y, Mohnen D, Gelineo-Albersheim I, Xu Y, Hahn MG (2011a) Glycosyltransferases of the GT8 Family. In: Ulvskov P (ed) Annual plant reviews: plant polysaccharides, biosynthesis and bioengineering. Wiley-Blackwell, Oxford, UK, pp 167–212
Yin Y et al (2011b) Evolution of plant nucleotide-sugar interconversion enzymes. PLoS One 6(11):e27995
Yin Y et al (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40(Web Server issue):W445–W4451
Yokoyama R, Nishitani K (2004) Genomic basis for cell-wall diversity in plants. A comparative approach to gene families in rice and Arabidopsis. Plant Cell Physiol 45(9):1111–1121
Yong W et al (2005) Genomics of plant cell wall biogenesis. Planta 221(6):747–751
York WS, O'Neill MA (2008) Biochemical control of xylan biosynthesis – which end is up? Curr Opin Plant Biol 11(3):258–265
Zhao QA et al (2010) Syringyl lignin biosynthesis is directly regulated by a secondary cell wall master switch. Proc Natl Acad Sci U S A 107(32):14496–14501
Zhong R, Ye ZH (2009) Transcriptional regulation of lignin biosynthesis. Plant Signal Behav 4(11):1028–1034

Chapter 9
Metagenomics of Plant–Microbe Interactions

Riccardo Rosselli and Andrea Squartini

The presence of plants on emerged land has been a main landscape-changing milestone in earth history. Their passage from planktonic lifestyle to that of rooted individuals stemming into air, along with the evolutionary shift to pluricellular forms, is a conquest that owes its foundation to a deal with microbes. The possibility of mineral nutrition through the access to growth-limiting resources, which could be scarce or in insoluble or unreachable forms, requires a major endeavor of microbial mediation to assist plants in direct and indirect ways. Likewise the possibility for a plant to survive, persist, or recur into an environment across seasons and years depends on the onset of a system that can recycle its residues and process the periodically deposited litter of perennial species. Organic matter decomposition with its main nutritional elements returning in mineral form, to the benefit of the vegetation is a lifelong ecosystem service in which microorganisms are the leading agents. Also the process of soil formation, starting from bare bedrock and working its way through centuries of slow pedogenesis is in itself an action that calls for a concerted interactivity unfolded by events led by microorganisms and plants. The former being responsible of several key actions. These include: initial colonization with autotrophic forms, rocky minerals solubilization via acidification; formation of crevices leading to a progressive increase of the surface/volume ratio in the mineral particles; mixing of organic matter, consisting in the microbe debris themselves, to the rocky portions of ground. The latter is a major step which introduces an altered geometry that breaks the ordered compactness of the clay minerals and creates the micropore-based environment that characterizes a mature underground habitat.

R. Rosselli, Ph.D.
Department of Biology, University of Padova, Via Ugo Bassi 58/b, Padova, Italy

A. Squartini, Ph.D. (✉)
Department of Agronomy Animals, Food, Natural Resources and Environment, DAFNAE, Viale dell'Università 16, Legnaro 35020, PD, Italy
e-mail: squart@unipd.it

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_9

Furthermore, organic matter and fungal mycelia that wrap particles are the keys for building the aggregate structure of soil that makes it hospitable for plants and more stable against erosion. Plants, which consequently find a substratum suitable for holding water, nutrients and anchoring themselves, in turn start to fuel the ground with a substantial portion of the organic carbon fixed by photosynthesis, which is exuded through roots and serves as a life-nourishing flow giving rise to the rhizosphere. This terms refer to the zone of influence of roots where microorganisms abund and multiply at rates several fold higher than those occurring in the mineral bulk soil. The rhizosphere is in this respect like a river running through a desert and it is the ground of many mutual benefits exchanged between plants and microbes, including growth-promoting activities, nutrient scavenging, solubilization and delivery, protection from pathogens, production of plant hormone-like compounds, and biostimulants of microbial origin. The presence of microbes can extend from soil into the plant as several species are able to be internalized as endophytes and invade the plant without pathogenic outcomes, conveying beneficial actions right from the inside of plant tissues.

As outlined above soil itself is a product of mutual actions between plants and microorganisms. Understanding its complexity requires that it is regarded not as an abstract entity but rather as one of the ends of the soil-rhizosphere-plant continuum, throughout which the microorganisms are actively distributed. Therefore soil should be seen as part of the greater plant–microbe interactome. The functioning of the terrestrial ecosystem and its productivity depend in turn on the metabolic activities and interactions happening in the soil, that we can consider in this respect, as the Earth's widest live interface.

The organisms dwelling in soil are very numerous. In a single gram, about a billion of live bacterial cells are typically present, and they belong, according to conservative estimates, to thousands of different species (Torsvik et al. 2002). Most of these are still unknown to biology due to the diffuse phenomenon of bacterial non-culturability in vitro which concerns over 99 % of environmental bacteria. This drawback imposes the use of methods, independent from culturing and isolation, to study soil biota in situ and characterize them by indirect means. Bacterial biomass ranges from 30 at 500 g/m^2. They share the soil environment with the other living members of groups as fungi (from 60 to 100 g/m^2), algae (0.5–10 g/m^2), protozoa (5–20 g/m^2) nematodes (0.1–0.3 g/m^2), arthropods (0.2–0.5 g/m^2), and anellids (0.5–200 g/m^2) (André et al. 2001; Lavelle and Sapin 2002).

As commented above, microbial life in soil contributes to the genesis and structuring of the soil itself and profoundly affects its attitude to host and support plants. Both productivity and health of the cultivated plants as well as the distribution of natural vegetation are conditioned by interactions with ground microbiota, with the most significant correlation being the presence or absence of symbiotic or pathogenic microorganisms (Klironomos 2002).

Besides their key role in recycling and degradation, soil microbes take active part in primary productivity both as photosynthetic unicellular entities and as light-independent chemolithotrophic species. Their ubiquity, versatility and survival strategies make them a very successful biological corporation driving both the

development and the stability of our physical environment; suffice to say that the total biomass of microbial cells on planet earth has been estimated equal to that of plants (10^{12} tons).

Despite its centrality, the soil, even the one nearest to our feet, largely remains for many aspects an unexplored environment. Its very high level of biological diversity and the consequent physio-metabolic complexity make this substratum a place whose knowledge requires new techniques with a correspondingly high power of resolution. The metagenomic approach, enabling to read, order, and compare very high numbers of extractable DNA sequences from the system, offers a route that, thanks to the recent technological development, has today become possible.

As a matter of fact the highly inhabited milieu that soil represents is at the same time the background from which plants derive their living and a complex biological community where the challenges of high throughput-assisted metagenomics can be put on trial in an ideal way.

It can be added that microbiology in general is a field that, most of any other branch of life, calls for a culture-independent approach since, as recalled above, the majority of bacterial species (>99 %) are not culturable on plates (Staley and Konopka 1985). This translates into an anomaly into the numbers of known species in comparison with botany and zoology, with particular respect to insects or to arthropods in general. In this respect the coming of age of metagenomics has offered the tool to fill this main gap in our biological knowledge, and to verify whether the known number of microbial species does represent the tip of a large iceberg or whether a different model of distribution (and/or a different species boundary definition) need to be applied to understand and describe earth microbiology.

Whichever the true structure of the unseen community be, and whatever the iceberg size, the techniques that are arising and improving, especially thanks to the NGS sequencing platforms, enable to verify many of the standing hypotheses and the first results have started to appear.

In covering the topic of metagenomics applied to plant and soil environmental microbiology some issues need to be clarified that apply to the correspondence between terminology and approach. First, as regards culture-independent DNA-based taxonomical studies of the members occurring in natural microbial communities, these studies have started in the 1980s, long before the availability of NGS machines (Olsen et al. 1986). Stemming from the path opened by Carl Woese in demonstrating the usefulness of the 16S rRNA sequence as universal comparative meter (Woese et al. 1975; Woese 1987), capitalizing the feasibility of PCR-based amplification of the corresponding 16s gene pools, a countless number of studies have addressed the composition of bacterial assemblages from the most diverse environments. Strictly speaking, this kind of approach, being directed to a single gene would however comply to the definition of metagenetics rather than of metagenomics and it was carried out in most cases through Sanger-sequencing of cloned amplicon libraries. As the present essay pertains to matter relating to NGS era, that vast body of data will not be reviewed here but it is to be acknowledged as the background reference that placed the basis from which we start our comprehension of microbial diversity and its interaction with our planet environments.

Focusing on the theme of this chapter, plants are a particularly important niche for microorganisms as these can be nourished by their organic compounds at different levels. First, in the soil surrounding the plant roots, its effect is felt as the rhizosphere; at closer range, the plant surface is a direct interface that unfolds underground as the rhizoplane and above on stems and leaves giving rise to the phyllosphere. The latter, consisting in the leaves surface, is characterized by harsh conditions of low nutrient outflow, direct sun exposure with high UV radiation impact, desiccating winds, but it is nevertheless a heavily colonized and extremely vast habitat (Yang et al. 2001; Lindow and Leveau 2002). Moving further to the internal parts of plants, both along vascular byways as well as right inside the tissues these zones are also home for an array of microbial species, the endophytes, that take advantage of such privileged location (Bacon and White 2000). The invasion of a host plant is in fact not only a matter of interaction with pathogenic species, but very often the outcome of a compromise that can span from apparently neutral or commensal attitudes (Berlec 2012), to various degrees of mutualism (Masson-Boivin et al. 2009). Pathogenic bacteria and fungi are however among the most studied as their economic damage to crops has called for highly funded research and extensive studies on their strategies of interaction with plants (Alvarez 2004). Some beneficial microorganisms appear to be very tightly associated with their host plants. Their persistence over generations within them is ensured by mechanisms of transmission to the offsprings (Zilber-Rosenberg and Rosenberg 2008). Such views have also led to the proposal that evolution of plants is correlated and concerted with that of their microbiomes, leading to the definition of the hologenome theory (Rosenberg et al. 2010).

There are several ways upon which microorganisms associated with plants can enact (Compant et al. 2010). These can span from plant-hormone production, to the biocontrol against plant pathogens, to drought stress-tolerance, to the solubilization of soil nutrients as phosphorus or iron. In the latter case the process unfolds via siderophores production (Wu et al. 2009). In the case of fungi, a number of species engage the mycorrhizal symbiosis which concerns the vast majority of land plants and confers to these an extended capability of exploring soil resources as limiting nutrients and water (Bonfante and Genre 2008). The potential of beneficial microorganisms in improving crop productivity has been exploited in the production and marketing of different types of agricultural inoculants, mostly as fertilizers or bioprotectors (Tikhonovich and Provorov 2011).

With the development of high throughput DNA sequencing techniques all these plant-microbe interactive scenarios have become approachable with a novel and deeper level of resolution (Handelsman 2004; Shendure and Ji 2008; MacLean et al. 2009). This should virtually overcome, at once, problems related to species unculturability, PCR-based biases in representing true species proportions within a complex community, rare species detection in species-rich environments (Tyson et al 2004; Tringe et al. 2005; Sogin et al. 2006).

In parallel to NGS principles and hardware, a corresponding line of bioinformatical solutions and user-friendly software tools is in parallel unfolding to match the massive data generated (Caporaso et al. 2010).

Studying microorganisms in association with plants can pose in first instance some technical challenges, for example when dealing with endophytes, one of the problems encountered when constructing a metagenomic library is the overwhelming excess of the host plant material (and DNA) with respect to the microbial one. Therefore different strategies can be chosen to enrich the microbial target either with detergents and salts (Wang et al. 2008) or with selective centrifugation to separate bacteria from plant cells (Sessitsch et al. 2012).

As regards the overall goals of the research, one possible question could be the assessment of which microbial species are associated with a particular plant or with a given part of it. In other words how hosts shape up and affect the community structure of their microbial partners. Other lines of research can focus on the effects of external challenges or stresses on the microbial community composition that can be found within or around a certain plant or vegetation.

A study on tree phyllosphere applied bar-coded pyrosequencing to survey bacterial identities on the leaves of 56 tree species across different locations (Redford et al. 2010). Results indicated a high degree of individual variability; conserved community patterns, coherent with the tree species, were detected, also irrespective of geographic distances and of surrounding atmosphere microbial composition. These data, obtained thanks to the high throughput of NGS machines, point out an important congruence between the phylogenies of trees and those of their phyllospheric microbiota. A certain difference stands out in this respect between the results of studies on the phyllosphere, whose microbial community is more determined by the host, and those on the rhizosphere where soil has a more pronounced effect and a plant recruits bacteria through the slow substrate outflow from roots. In phyllospheres instead the selection process appears to be more rapid and is exerted right on the plant leaf surface, more on a compatibility basis than by substrate-driven mechanisms (Bulgarelli et al. 2013)

The situation of the plant as major determinant of phyllospheric community is nevertheless not to be generalized as other studies on *Tamarix* shrubs belonging to three different species (Finkel et al. 2011, 2012), using 16S/18S rRNA tag pyrosequencing, examined over 200,000 sequences of the 16S V6 and 18S V9 hypervariable regions and revealed a diverse community, with 788 bacterial and 64 eukaryotic genera. The analysis pointed out that the geographical issue and local surroundings are important determinants of the bacterial assemblages composition. In a different study poplar endophytes and rhizospheric microorganisms were investigated by community profiling using 454 pyrosequencing with separate primers for the V4 region for bacterial 16S rRNA and the D1/D2 region for fungal 28S rRNA genes (Gottel et al. 2011). Results indicated the distinctness of the two compartments composition. The rhizosphere dominant bacteria resulted Acidobacteria (31 %) and Alphaproteobacteria (30 %), while inside the plant roots the scores were led by Gammaproteobacteria (54 %), in large part represented by a *Pseudomonas* sp. The diversity of bacterial communities colonizing the oak rhizosphere and the surrounding soil was studied upon analyzing, by 454 pyrosequencing, over 300,000 16S rDNA amplicons (Uroz et al. 2010). Acidobacteria and Proteobacteria resulted dominant but the relative proportions of the overall phyla and of the Proteobacteria classes clearly differentiated the rhizosphere from the bulk soil environment.

A pyrosequencing-based analysis of the V2–V3 16S rRNA gene region was undertaken to identify fluctuations in bacterial diversity in nine forest and nine grassland soils (Nacke et al. 2011). The dataset encompassed 598,962 sequences. Bacterial diversity resulted to be more phylum-rich in grassland soils than in forest soils: Acidobacteria, Proteobacteria, and Actinobacteria being the most prominent taxa. Community structure had moreover significant differences between beech and spruce forest soils. Besides the dependence on vegetation, bacterial diversity was correlated with soil pH, but not particularly with land management type nor other soil properties.

A further study was devoted to monitoring the responses of rhizospheric bacteria to different low molecular weight carbon substrates as glucose, serine or citric acid via bar-coded pyrosequencing of the 16S rRNA gene (Eilers et al. 2010). All three substrates had effect in 24h which mostly consisted in a stimulation of Betaproteobacteria, Gammaproteobacteria, and Actinobacteria. Citric acid was in particular active on the former group with effects displaying a two- to fivefold increase.

Some authors have addressed the question of how the domestication of plants and the selection of modern crops could have been shaped by the underlying relationships with microbes, and investigated this topic via 16S rRNA taxonomic microarray, targeting 19 bacterial phyla on five main genetic inbred maize lines rhizospheres (Bouffaud et al. 2012). Results indicated a correlation between the bacterial assemblages and the maize genetic structure, mostly explained by Betaproteobacteria of the genus *Burkholderia*.

In a study on the rhizosphere microorganisms of six potato varieties at three developmental stages in two types of soils, the authors used DNA-based pyrosequencing and screened over 350,000 sequences (İnceoğlu et al. 2011). While the rhizosphere samples were markedly different from the bulk soil ones and the dominant members were mostly represented by Actinobacteria, and Alphaproteobacteria, no clear host genotype selection was evident, with only some effect at the plants youngest stage.

However, when evaluating the results of all these studies it must be borne in mind that when dealing with techniques such as 16S rRNA sequencing as a tool to extract taxonomical identity information, this gene is not carrying a suitable variability below the species level to enable the assessment of subspecies or strain-related differences. As a consequence, when the relationships between host plant genotype and specific microorganims are expressed at these levels, different approaches as whole metagenome sequencing and comparison can be envisaged more suited to reveal possible host specificity issues.

The identity of bacteria present on tomato fruits surface in relation to different irrigation water sources was assessed by 454 pyrosequencing (Telias et al. 2011). While there were significant differences in bacterial assemblages between surface water (displaying higher species richness with abundant gram-positives) and groundwater (less diverse and dominated by proteobacteria), the fruit surface ended up colonized by a community that was not significantly distinct. Also this experience shows the profound effect of a plant habitat in specifically selecting and enriching some precise members of the bacterial array that are able to become established in that particular niche.

Pyrosequencing-based analysis of the V1–V2 16S rRNA gene region coupled to Community level physiological profiling (CLPP) were used to analyze bacterial diversity from soils of chemically cultivated land, organically cultivated land, and fallow grass land, after 16 years of such managements (Chaudhry et al. 2012). The most abundant phyla were Proteobacteria (29.8 %), Acidobacteria (22.6 %), Actinobacteria (11.1 %), and Bacteroidetes (4.7 %), Proteobacteria, Bacteroidetes, and Gemmatimonadetes were more abundant in organically cropped soils while Actinobacteria abounded in chemically treated ones, and Acidobacteria in fallow soils.

Simulation of climate change by altering CO_2 concentrations was tested in its effects on nitrifying archeal and bacterial community compositions in soybean rhizospheres, which were assessed by 454 sequencing on amplicon libraries of 16S rDNA and on the nitrification gene *amoA* (Nelson et al. 2010). The archeal phyla were found affected in their relative proportions but the copy number of the nitrification gene was not.

In some studies, the presence of genes linked to key microbial proteins as phototrophic pigments were investigated (Atamna-Ismaeel et al. 2012). 454-pyrosequencing-generated metagenome data from the phyllospheres of five plants: tamarisk (*Tamarix nilotica*), soybean (*Glycine max*), Thale cress (*Arabidopsis thaliana*), white clover (*Trifolium repens*), and rice (*Oryza sativa*). The data revealed for the first time the presence of bacterial rhodopsins outside of aquatic habitats and showed that microbial and plant photosynthesis co-occur in such habitat. Rhodopsin is a pigment similar to those involved in animal vision. Its spectrum of light absorbance is different from that of chlorophyll implying an adapted noncompeting instance of coexistence between plants and the bacteria on their leaves.

The host plant factor resulted dominant in shaping the structure of the bacterial community borne on its leaves. In a study on lettuce belonging to different cultivars, even if those were grown in the same field, the foliar bacterial communities were distinct (Hunter et al. 2010), leading to the hypothesis that leaf texture and leaching of specific metabolites were the basis of these specificities. A further understanding of the plant genotype role in selecting specific microorganisms came from a study on the phyllosphere of recombinant inbred corn varieties (Balint-Kurti et al 2010). QTL mapping revealed the existence of chromosomal loci that control epiphytic microbiota. These coincide with the determinants that underlie the susceptibility to a fungus causing the Southern leaf blight disease.

Some plants, as those of the Leguminosae family can engage at the same time in different symbioses with nitrogen-fixing bacteria and with mycorrhizal fungi. Many of the legumes are important food crops and the analysis of their relationships with microbes can help to foster their productivity by improving field inoculation technologies. The availability of soybean mutants defective in the symbiotic nodule formation has allowed an experiment in which plant shoot colonization by different bacteria has been analyzed in relation to plant genotype and to nitrogen fertilization levels (Ikeda et al. 2010). The study reported that both the symbiotic attitude and the exogenous nitrogen supplementation can affect internal colonization by bacteria other than the *Bradyrhizobium* symbionts. The recurring colonizers included *Methylobacterium* and *Aurantimonas*.

Studies reviewing patterns of bacterial dynamics in their interactions with plants have put in evidence some key trends (Bulgarelli et al. 2013). Data indicate a dual step selection process which initially leads to a differentiation of rhizosphere biota from those of bulk soil. Subsequently the rhizodeposition of organic compounds leached from roots further selects some taxa also via gene for gene catabolic interactions on specific trophic compounds. A defined guild of microbes is eventually gaining access into plant tissues and becoming endophytic also by virtue of a biochemical mutualism.

In some cases the studies of plant–microbe interactions have not been targeted at situations occurring in nature but to vegetables in the food processing and marketing contexts. From this standpoint the dynamics of phyllospheric bacteria in relation to storage of spinach at different refrigeration temperatures was investigated by pyrosequencing of 16S rRNA gene amplicons (Lopez-Velasco et al. 2011). Notwithstanding the nature of horticultural crops and the man-managed artificial environment, a remarkable degree of initial diversity was noticed in the microorganisms present on leaves. The analysis revealed over 1,000 operational technical units (OTUs) of which 75 % were associated with previously undescribed taxa. At least 11 prokaryotic phyla were represented, among which the most abundant cases belonged to Proteobacteria and Firmicutes. The refrigeration and packaging procedures showed a rapid decline of the richness with a drop to only 5 phyla after one day. Such conditions promoted a shift in the community with the fast rise of Gammaproteobacteria, as *Pseudomonas* spp. and different Enterobacteriaceae which became the most abundant taxa after 15 days at either 4 or 10 °C. However, within the family, different genera displayed individual cold-sensitivity as the growth of *Escherichia coli* was inhibited at the lower temperature. The study demonstrated a relevant application of NGS techniques to the field of food safety and quality control. Within the increasing awareness of these possibilities it has been postulated that phyllospheric communities could be purposely modulated by plant breeding or by selected agrochemical treatments in order to enhance fresh produce safety (Newton et al. 2010).

Another study addressed a further agronomical-level case by analyzing leaf bacterial communities on lettuce (Rastogi et al. 2012). The analysis also addressed the changes due to time and space by following, on the same crop, geographical and seasonal transects in California and Arizona. Besides the pyrosequencing of 16S rRNA gene amplicons, the project involved also the quantitative assessment of phyllospheric bacteria and of their culturability on plates. The total bacterial populations on leaves ranged between 10^5 and 10^6 cells per gram of tissue and the fraction of culturable colonies was ten times lower in summer than in spring and a hundred times lower in winter. Sequencing of the small subunit ribosomal rRNA gene indicated as most abundant groups the Proteobacteria, Firmicutes, Bacteroidetes, and Actinobacteria at phylum level, and within these *Pseudomonas, Bacillus, Massilia, Arthrobacter*, and *Pantoea* at genus level. The authors considered these taxa as a representative core of the bacteria associating with cultivated lettuce irrespective of season or localities. The analysis allowed also to find indicator species that correlated with potential pathological conditions. In this respect the

presence of some known pathogens of this crop, such as *Xanthomonas campestris* pv. *vitians*, was in positive correlation with the presence of bacteria from the genus *Alkanindiges*, and in negative correlation with the genera *Bacillus, Erwinia*, and *Pantoea*. As regards seasonal variations, the summer sampling scored higher presences of Enterobacteriaceae and culturable coliforms in comparison with the winter situations. Geographically there were trends that correlated with the distance but were not affected by the cultivar of lettuce analyzed. Also the effect on an impactful event, consisting in a dust storm, was evaluated in that survey and its effect in alteration of the bacterial assemblages was quantified.

The emerging view of hitherto unnoticed members of the microbial entourage of a plant as key elements for disease prevention and overall fitness is becoming clearer also thanks to metagenomics (Babalola 2010; Bulgarelli et al. 2013). An example comes from the case of *Sphingomonas* that was demonstrated to interfere with the growth of pathogenic *Pseudomonas syringae* pv. *syringae* in the phyllosphere of *A. thaliana* (Innerebner et al. 2011).

Among the questions that could be addressed via metagenomics stands the issue on what is the environment from which the phyllosphere microorganisms are recruited. Contrary to some expectations, soil was not the primary source as shown by pyrosequencing of phyllosphere biota with those of the surrounding soil, which resulted to share only 0.5 % of the sequences (Kim et al. 2012). But if soil was not recognized as source for epiphytic microbes, air was not either, as suggested by studies comparing airborne community with that stabilizing on leaves (Vokou et al. 2012). A parallel study on the model plant *A. thaliana* yielded Actinomycetales and Actinoplanes as most abundant in rhizosphere bacteria whereas the phyllosphere was dominated by *Pseudomonas* (Bodenhausen et al. 2013).

The comparative genomics of sequenced bacterial epiphytes as that of *Pantoea agglomerans* 299R (Remus-Emsermann et al. 2013) is starting to reveal the set of genes that enable successful persistence in the rhizosphere. These include specific sugar utilization, osmoprotection, and DNA repair systems to cope with UV damage. Also the formation of a capsule with copious extracellular polysaccharides (EPS) results a common and crucial trait to form cell aggregates and prevent desiccation and osmotic stress (Monier and Lindow 2004).

When entering a plant, bacteria encounter pros and cons at each level from surface to the inner apoplast. The outer stages expose them to UV radiation and drought while inside the plant they need to come to terms with defense compounds. Transcriptional profiling of the leaf pathogen *P. syringae* which can access both microenvironments, showed that genes such as those for motility, chemosensing, phosphate and sulfur uptake, and indole metabolism to derive tryptophan, were preferentially expressed on the surface location rather than in the apoplast (Yu et al. 2013). On the contrary genes involved in the metabolism and transport of gamma-aminobutyric acid, production of secondary metabolites, and phytotoxins such as syringomycin, and syringopeptin were turned on when the bacterium had entered inside the plant.

Besides bacteria, plant surface can be colonized by fungi although in lower numbers. A high- throughput pyrosequencing analysis of fungal internal transcribed

spacer 1 (ITS1) was carried out to detect the fungi on the phyllosphere of oak (Jumpponen and Jones 2009) and the rural environment was compared to urban management areas. In the former, prevailing taxa included *Ramularia, Stenella, Dioszegia, Devriesia, Mycosphaerella, Paraphaeosphaeria, Phaeosphaeria,* and *Sphaeceloma*, while in the latter the community was led by *Aureobasidium, Davidiella, Didymella*, and *Microsphaeropsis*. The study highlighted the importance of land use in fungal colonization of the host as more important than plant species. It can be added that the nutrient content (in particular concerning nitrogen and sulfur) could be a discriminant feature between urban and forest settings leading to uneven contents within the same plant species and explaining phenotypic outcomes such as the associated biota.

Further exploring plant epiphytic fungi by 454 pyrosequencing of the ITS1 region, those thriving on leaves of *Fagus sylvatica*, the European beech, were studied, distinguishing also different parts of the plants (Cordier et al. 2012). A large presence of cosmopolitan fungi as *Taphrina, Lalaria*, and *Woollsia* was recorded and the community variability was mostly evident at the small spatial scale represented by the leaves. The host genotype was a strong determinant of community composition as the genetic distance between beech trees imposed corresponding differences in fungal assemblages. Again by analyzing balsam poplar phyllosphere by ITS pyrosequencing (Bálint et al. 2013) a majority of plant genotype-specific fungal taxa was pointed out.

The microbiology of the phyllospere was studied also by combining more -omics approaches and conjugating metagenomics with metaproteomics into metaproteogenomics (Delmotte et al. 2009). Soybean, clover, and *A. thaliana* leaves were studied, and recurring distinct bacterial communities were identified. The proteomic approach enabled to trace physiological traits as characteristic such as enzymes for the use of one-carbon plant compounds as methanol, which correlated with the presence of *Methylobacterium,* and TonB-dependent receptors which were linked to the presence of *Sphingomonas*. This example shows the usefulness of coupling protein sequencing via mass spectrometry with their identification on the sequenced metagenome from the same sampling.

A further work on metaproteogenomics addressed the comparison of leaf and root microbiota in rice, searching for distinct metabolic processes that could characterize the two compartments. DNA was analyzed by using in parallel a 16S rRNA amplicons sequencing approach and a whole metegenome 454 sequencing (Knief et al. 2012). At phylum level the dominant divisions in the phyllosphere were Actinobacteira (38 %) and Alphaproteobacteria (35 %). In the rhizosphere Alpha-, Beta-, and Deltaproteobacteria together accounted for 10 % of the cases while all the other phyla had less than 5 % and included Actinobacteria, Firmicutes, Gammaproteobacteria, and Deinococcus-Thermus. The percentage on unknown taxa, not classifiable at the cutoff identity imposed, was much higher in the rhizosphere and reached 40 %. The phyllosphere was characterized by genera as *Rhizobium, Methylobacterium*, and *Microbacterium.* The proteomics part of the experiment enabled to cover more than 4,600 identified proteins, whose analysis revealed enzymes for methylotrophy such as methanol dehydrogenase, formaldehyde

activating enzyme, which comply with the phenotype of Methylobacterium. In the rhizosphere instead, genes for the methanogenesis and methanotrophy were found. Moreover the transport processes and the stress responses were more pronounced in the phyllosphere. In both sectors of the plant, genes for nitrogen fixation were found, but as regards their expression that occurred only in the rhizosphere. This example shows the investigative advantages of combining metagenomics with metaproteomics to achieve a deeper level of information on environmental activities related to microorganims.

A different way to address the interactive assets of bacteria towards plants by NGS methodologies is achieved by sequencing different isolates of the same species and comparing their sequences to enucleate the conserved core of genes which are supposedly relevant for their ecology. This kind of reverse metagenomics has been applied to the study of *Pseudomonas fluorescens* (Loper et al. 2012) whereby seven strains were sequenced. They differ in their capability of biocontrol vs plant pathogens and some can also interact with insects. The approach allowed to distinguish the isolates into three subclades by using multilocus sequence analysis. The pan-genome of *P. fluorescens* species resulted very large as it accounted for about 54 % of the pan-genome of the whole *Pseudomonas* genus. This was distinguished from the core genome which represented just 45–52 % of the genome of any individual strain. The study also allowed to discover novel genes including those for some siderophores, antibiotics, bacteriocins, secretion systems of different types and toxins active against insects. This approach also enabled a subsequent phylogenetic inference work by examining the distribution of repetitive palindromic extragenic (REP) elements in relation to that of mobile genetic elements as insertion sequences and transposons.

Other methodologies that have been successfully used to assist the detection of plant-associated microorganisms include the Phylochip-based metagenomics (Mendes et al. 2011). The principle is based on a microarray-hybridization on chip, that supports a defined but vast series of species-specific primers. The approach was tested on disease-suppressive soils, i.e. those in which plants tend to be immune from a number of microbial pathogens infections that would cause severe damages to them if grown elsewhere. The basis of suppressiveness is thought to be due to the presence of antagonistic microorganisms as soil sterilization by heath can cancel the suppressive properties. Aiming at defining also whether the presence of specific microbiota could explain such protective effect, the phylochip was applied to rhizosphere of sugarbeet seedlings, infecting those with the fungal pathogen *Rhizoctonia solani*, a major pest for many agricultural crops, and growing them in disease-suppressive or disease-conducive soil. The analysis led to the definition of 33.346 OTUs of bacterial and archeal nature, which reveals, in both types of soil, a high species richness, also compared to previous studies. Dominant phyla were Proteobacteria (39 %) and Firmicutes (20 %). A large fraction of unclassified cases (16 %) was also present. To verify a difference in the two cases, the Bray-Curtis dissimilarity matrix analysis was calculated on the relative abundance of the bacterial and archaeal OTUs identified. The rhizosphere communities of the suppressive and conducive soils consequently showed their differences. A number of bacterial

taxa were ascribed to suppressiveness. These included the families of Pseudomonadaceae, Burkholderiaceae, Xanthomonadales and Lactobacilaceae and the phylum Actinobacteria. The former family in particular featured a *Pseudomonas* producing a chlorinated lipopeptide encoded by a non-ribosomal peptide synthase, which was shown to be a major determinant of disease-prevention. Other taxa however appear to concur to the overall phenomenon, indicating the synergistic action of a microbial consortium as the basis for the full effects of these soils.

The involvement of *Pseudomonas* has a parallel in a different disease control situation as other strains of this species are known to produce the antibiotic compound 2,4-diacetylphloroglucinol (2,4-DAPG) which suppresses the fungal pathogen, *Gaeumannomyces graminis*, in soil thereby defined as suppressive of the "take-all disease" which this fungus causes to cereals (Raaijmakers and Weller 1998). In the phylochip-based study (Mendes et al. 2011) upon isolating different *Pseudomonas* from the suppressive soil, despite the fact that most were able to colonize the rhizosphere, only one of them was able to confer the suppressiveness phenotype.

A different perspective, still on the plant disease interactive scenarios, addressed a different question: How is the rhizosphere community altered by a plant pathogen? The study involved a combined approach of microarray chips (Geochip 3.0). Clone library sequencing and taxon/group-specific quantitative real-time PCR applied to citrus rhizosphere in plants affected by the severe disease known as huanglongbing (HLB), caused by the bacterium *Candidatus* Liberibacter asiaticus (Trivedi et al. 2012). The phylum Proteobacteria, with genera of ascertained rhizosphere colonization attitude, was the most abundant in healthy rhizospheres while phyla as Acidobacteria, Actinobacteria, and Firmicutes, which were dominant in the bulk soil resulted more represented in soil infected with the HLB disease agent.

The study also addressed which genes and functions are affected. The phylochip analysis showed that HLB disease impacted on different bacterial guilds; a number of genes responsible of known ecological processes as nitrogen cycling, carbon fixation, phosphorus utilization, metal homeostasis and resistance, resulted differentially present, displaying higher scores in healthy rhizospheres of citrus. The microbial community in the diseased soil appeared to have undergone a shift from the use of readily degradable forms of carbon to those that are rather recalcitrant.

Among the recent studies that have explored the diversity of plant-associated microbes, a paradigmatic report is represented by two papers which appeared on the same issue of Nature focusing on the model species *A. thaliana* (Lundberg et al. 2012; Bulgarelli et al. 2012).

The former describes the high-resolution 16S amplicon pyrosequencing comparing bulk soil, rhizosphere, and endophytic microorganisms, upon taking into account over 600 individual plants, in two soil types. The project tested the hypothesis that the bacteria ending up in the different compartments would mostly depend on host genotype. 778 microbial OTUs were individuated. 12 of these demonstrated plant genotype-related quantitative enrichment within the endophytic domain. Results showed that, while soil has an effect on determining the communities composition, the endophytic inner core of bacteria shows a conspicuous degree of overlapping and is characterized by low-complexity assemblages, enriched in Actinobacteria

and given families of Proteobacteria. Plant developmental stage and genotype also account for a part of the observed variability. With the exception of some transient dwellers, the true endophytes of this plant, which as many cruciferae lacks fungal mycorrhizal associations, appear to be a well-defined group of conserved taxa with either symbiotic or potentially pathogenic attitudes.

In the parallel study (Bulgarelli et al. 2012) two ecotypes of *A. thaliana* in two soils (of clay and sandy type) were examined by pyrosequencing of the 16S ribosomal RNA gene PCR amplicons, analyzing the variable V5–V6 regions. Plants appeared to be mostly invaded by Proteobacteria, Bacteroidetes, and Actinobacteria, with each of these phyla being represented by a dominant class or family. The effect of soil type in determining the members of the endophytic community was appreciable supporting the view of a soil origin for plant endophytes. The host plant genotype appeared to have a certain effect in determining which bacterial ribotype profiles were found within the endophytes community. In the same study the authors also tested the microbe-enriching capabilities of inert wooden material inserted in the same soil, in order to derive information on the effect of metabolically inactive lignocellulosic material. Interestingly the resulting colonizers had 40 % of the taxa in common with the *Arabidopsis* root-enriched ones, with a high representation of Betaproteobacteria. On the contrary Actinobacteria were underrepresented in these wood-enriched groups, being more neatly associated with the live root-selected community.

As regards endophytic taxa, Tag Encoded FLX amplicon pyrosequencing (TEFAP) was used to detect pCR-amplifiable ribosomal genes of bacteria and fungi in a work on micropropagated *Atriplex* plants (Lucero et al. 2011). A series of different primer pairs were tested and their performances compared. In this approach the authors also performed a comparison with culturable colonies from the plant tissues under examination and a parallel observation of the material using light, electron, and confocal microscopy. The molecular analyses revealed 7 bacterial and 17 ascomycete taxa in *A. canescens*, and 5 bacterial taxa in *A. torreyi*. The use of different primers revealed differences in the results pointing out the problem of potential primer biases and preferential amplification of certain targets, which occurs even in a low complexity community. This evidence raises a general issue of the accurate representation of a community when studies rely on a PCR-dependent amplification step. The microscopy complementation of this study also indicated the presence of microbial cells in leaves and root tissues that had been regenerated aseptically. These and other clues suggested the hypothesis that endophytes of these plants are seedborne and transmitted to the progeny.

Another relevant piece of research focusing on the plant endophytism on a major crop, rice, reveals other clues on the identities of endophytic bacteria, on their lifestyle, and on proteins relevant for endophytism (Sessitsch et al. 2012). That work, not using a PCR-dependent strategy but targeting full metagenome sequencing on isolated endophytic bacterial cells, also followed an original approach of endophyte separation via centrifugation. This, likewise, involved a massive amount of field-grown rice plants biomass to be processed in order to purify enough bacterial DNA for efficient metagenomics. Protein putative functions were deduced from similarity analyses and suggested which traits and metabolic pathways are relevant

for endophytic life. These involved the presence of flagella, plant-polymer-degrading enzymes, protein secretion systems, iron uptake and storage, quorum sensing, and reactive oxygen species detoxification. As regards genes for the nitrogen transformation, interestingly evidences have arisen that the whole cycle could occur in the endorhizosphere, as genes for nitrogen fixation, nitrification and denitrification were present in the endophytic populations. Results extend the knowledge on endophytism and open interesting perspectives for its exploitation, regardless the non-culturability of the bacteria, for the improvement of plant productivity, stress resistance, biological control, and bioremediation.

A thorough application of sequencing techniques calls for the view from the transcriptional side of the gene expression row. In this respect studies have appeared that consider the whole transcriptome analysis applied to plant microbe interactions and concern in this case the host plant transcriptomics. This allows, for example, to scan gene expression at timely chosen moments after inoculation of a pathogen. It was the case of a study on peach leaves after infection by *Xanthomonas arboricola* pv. *pruni* (Socquet-Juglard et al. 2013). The bacterium is a threat for peach production causing necroses on fruits and leaves that have a severe impact on crop yield and commercialization-preventing damage. Little evidence was yet available on defense mechanism in peach trees. A whole transcriptome sequencing analysis scanned the differentially expressed genes at two time points, consisting in 2 and 12 h post infection, in comparison with non-inoculated control leaves. Out of the 19,781 known peach genes that were expressed in all time points, there were 23 and 263 that were expressed differentially at 2 and 12 h post infection. Within these, 82 and 40 % were up-regulated; and 18 and 60 % were down-regulated. Bioinformatical searches against the Gene Ontology database indicated the stress response genes particularly represented at 2 h, while cellular and metabolic processes genes were more involved in the second time point. The differential expression of genes related to known pathogen-associated molecular pattern (PAMP) receptors, of disease resistance genes (including several RPM1-like and pathogenesis related thaumatin-encoding genes) was particularly informative of the defense pathways that the plant adopted. The peach also showed response in genes for photosynthesis, citochromes, reorganization of cell wall and hormone signalling, along with other transcripts providing several field-advancing pieces of information in early stages of plant defense from bacterial pathogens.

A different study, also using a transcriptomics approach and focusing on plant pathogen interaction, showed that with massively parallel transcriptome sequencing one can study the fungus and the infected host in parallel and enucleate the arsenal of effectors that are used by the two interacting genomes (Stassen et al. 2012). The analysis was focused on lettuce downy mildew (*Bremia lactucae*) an oomycete using secreted proteins in its invasive interaction strategy. The effectors that are known to be translocated by the host belong to two classes: RXLR and Crinklers. The massively parallel sequencing was done on cDNA derived on one side from *B. lactucae* fungal spores, and on the other, from infected lettuce seedlings. More than 2.3 million 454 FLX reads were processed, assembling 59,618 contigs representing transcripts from both organisms. 19,663 of these belonged to *B. lactucae* as they

matched SOLiD genome sequences stemming from more than 270 million reads from spore DNA. Upon translating these into protein models there were 1,023 of these that suited the prediction of being part of a secretome that featured elicitins, necrosis, ethylene-inducing peptide 1-like proteins, glucanase inhibitors, and lectins, and was enriched in cysteine-rich proteins. There were likewise 78 candidate host translocated effectors complying with the RXLR or with the Crinkler types.

In essence, the advent of metagenomics and the availability of the NGS machines have initiated an age of important discoveries in several fields and in particular in that of environmental microbiology in which the soil and plant contexts occupy a prime position. The taxonomical complexity of microbial communities and their richness are nevertheless still posing robust challenges to investigations of this kind. The sensu stricto metagenomical approach would call for sequencing of the DNA totality. Notwithstanding the high throughput, when the goal is unravelling the identities of many different microbes hosted by a given habitat, studies tend to focus on defined genes, such as the 16S rRNA determinants rather than on the whole metagenome. Besides the terminology issue which would make this a metagenetic rather than metagenomic approach, some cautionary considerations ought to be made in this respect. Since the standard way of tackling the 16S analysis is a PCR-driven enrichment of the amplicon, making use of bona fide universal primers, a variable degree of primer bias, and uneven mismatches in the first PCR cycles can severely alter the true community proportions at the endpoint stage of the amplification reaction. Moreover, the matching primer assumption is a self-referencing condition that rests on an a priori knowledge which rests on previously observed taxa but might be incomplete as regards the remaining yet-to-find taxa. Possible novel species that could be just out of the boundaries of the conserved consensus could even not be amplified at all, and thence never discovered.

For this reason, it is auspicated that future developments of studies of mixed microbial assemblages could be carried out by means of PCR-independent approaches, one of these could be, for example, the direct 16S rRNA sequencing, which could at once give an unbiased picture of the community members, coupled to their levels of gene expression activity, and consequently reveal hitherto unknown taxa with higher polymorphism in their ribosomal genes. It can be foreseen that novel protocols will develop these concepts and improve our knowledge in the plant-microbe interactive scenario as well as in other environmental contexts.

References

Alvarez AM (2004) Integrated approaches for detection of plant pathogenic bacteria and diagnosis of bacterial diseases. Annu Rev Phytopathol 42:339–366

André H, Ducarme X, Anderson JM, Behan-Pelletier V, Crossley DA Jr, Koehler HH, Paoletti MG, Walter DE, Lebrun P (2001) Skilled eyes are needed to study soil's richness. Nature 409:761

Atamna-Ismaeel N, Finkel OM, Glaser F, Sharon I, Schneider R, Post AF, Spudich JL, von Mering C, Vorholt JA, Iluz D, Beja O, Belkin S (2012) Microbial rhodopsins on leaf surfaces of terrestrial plants. Environ Microbiol 14:140–146. doi:10.1111/j.1462-2920.2011.02554.x, Epub 2011 Sep 1

Babalola OO (2010) Beneficial bacteria of agricultural importance. Biotechnol Lett 32:1559–1570

Bacon CW, White J (2000) Microbial endophytes. CRC, New York

Bálint M, Tiffin P, Hallström B, O'Hara RB, Olson MS, Fankhauser JD, Piepenbring M, Schmitt I (2013) Host genotype shapes the foliar fungal microbiome of balsam poplar (*Populus balsamifera*). PLoS One 8:e53987

Balint-Kurti P, Simmons SJ, Blum JE, Ballaré CL, Stapleton A (2010) Maize leaf epiphytic bacteria diversity patterns are genetically correlated with resistance to fungal pathogen infection. Mol Plant Microbe Interact 23:473–484

Berlec A (2012) Novel techniques and findings in the study of plant microbiota: search for plant probiotics. Plant Sci 193–194:96–102

Bodenhausen N, Horton MW, Bergelson J (2013) Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. PLoS One 8(2):e56329. doi:10.1371/journal.pone.0056329

Bonfante P, Genre A (2008) Plants and arbuscular mycorrhizal fungi: an evolutionary-developmental perspective. Trends Plant Sci 13(9):492–498. doi:10.1016/j.tplants.2008.07.001, Epub 2008 Aug 12

Bouffaud ML, Kyselkova M, Gouesnard B, Grundmann G, Muller D, Moenne-Loccoz Y (2012) Is diversification history of maize influencing selection of soil bacteria by roots? Mol Ecol 21:195–206

Bulgarelli D, Rott M, Schlaeppi K, Loren V, van Themaat E, Ahmadinejad N, Assenza F, Rauf P, Huettel B, Reinhardt R, Schmelzer E, Peplies J, Gloeckner FO, Amann R, Eickhorst T, Schulze-Lefert P (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature 488:91–95. doi:10.1038/nature11336

Bulgarelli D, Schlaeppi K, Spaepen S, Ver Loren van Themaat E, Schulze-Lefert P (2013) Structure and functions of the bacterial microbiota of plants. Annu Rev Plant Biol 64:807–838

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336

Chaudhry V, Rehman A, Mishra A, Chauhan PS, Nautiyal CS (2012) Changes in bacterial community structure of agricultural land due to long-term organic and chemical amendments. Microb Ecol 64(2):450–460. doi:10.1007/s00248-012-0025-y, Epub 2012 Mar 15

Compant S, Clement C, Sessitsch A (2010) Plant growth-promoting bacteria in the rhizo- and endosphere of plants: their role, colonization, mechanisms involved and prospects for utilization. Soil Biol Biochem 42:669–678

Cordier T, Robin C, Capdevielle X, Fabreguettes O, Desprez-Loustau ML, Vacher C (2012) The composition of phyllosphere fungal assemblages of European beech (*Fagus sylvatica*) varies significantly along an elevation gradient. New Phytol 196:510–519

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C, Vorholt JA (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci U S A 106:16428–16433

Eilers KG, Lauber CL, Knight R, Fierer N (2010) Shifts in bacterial community structure associated with inputs of low molecular weight carbon compounds to soil. Soil Biol Biochem 42:896–903

Finkel OM, Burch AY, Lindow SE, Post AF, Belkin S (2011) Geographical location determines the population structure in phyllosphere microbial communities of a salt-excreting desert tree. Appl Environ Microbiol 77:7647–7655

Finkel OM, Burch AY, Elad T, Huse SM, Lindow SE, Post AF, Belkin S (2012) Distancedecay relationships partially determine diversity patterns of phyllosphere bacteria on Tamarix trees across the Sonoran desert. Appl Environ Microbiol 78:6187–6193

Gottel NR, Castro HF, Kerley M, Yang Z, Pelletier DA, Podar M, Karpinets T, Uberbacher E, Tuskan GA, Vilgalys R, Doktycz MJ, Schadt CW (2011) Distinct microbial communities

within the endosphere and rhizosphere of Populus deltoides roots across contrasting soil types. Appl Environ Microbiol 77:5934–5944
Handelsman J (2004) Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev 68:669–685
Hunter PJ, Hand P, Pink D, Whipps JM, Bending GD (2010) Both leaf properties and microbe-microbe interactions influence within-species variation in bacterial population diversity and structure in the lettuce (Lactuca species) phyllosphere. Appl Environ Microbiol 76:8117–8125
Ikeda S, Okubo T, Anda M, Nakashita H, Yasuda M, Sato S, Kaneko T, Tabata S, Eda S, Momiyama A, Terasawa K, Mitsui H, Minamisawa K (2010) Community- and genome-based views of plant-associated bacteria: plant–bacterial interactions in soybean and rice. Plant Cell Physiol 51:1398–1410
İnceoğlu Ö, Al-Soud WA, Salles JF, Semenov AV, van Elsas JD (2011) Comparative analysis of bacterial communities in a potato field as determined by pyrosequencing. PLoS One 6(8):e23321. doi:10.1371/journal.pone.0023321
Innerebner G, Knief C, Vorholt JA (2011) Protection of *Arabidopsis thaliana* against leafpathogenic *Pseudomonas syringae* by Sphingomonas strains in a controlled model system. Appl Environ Microbiol 77:3202–3210
Jumpponen A, Jones KL (2009) Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. New Phytol 184:438–448
Kim M, Singh D, Lai-Hoe A, Go R, Abdul Rahim R, Ainuddin AN, Chun J, Adams JM (2012) Distinctive phyllosphere bacterial communities in tropical trees. Microb Ecol 63:674–681
Klironomos JN (2002) Feedback with soil biota contributes to plant rarity. And invasiveness in communities. Nature 417:67–70
Knief C, Delmotte N, Chaffron S, Stark M, Innerebner G, Wassmann R, von Mering C, Vorholt JA (2012) Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. ISME J 6:1378–1390
Lavelle P, Sapin AV (2002) Soil ecology. Kluwer, Amsterdam
Lindow SE, Leveau JH (2002) Phyllosphere microbiology. Curr Opin Biotechnol 13:238–243
Loper JE, Hassan KA, Mavrodi DV, Davis EW 2nd, Lim CK, Shaffer BT, Elbourne LD, Stockwell VO, Hartney SL, Breakwell K, Henkels MD, Tetu SG, Rangel LI, Kidarsa TA, Wilson NL, van de Mortel JE, Song C, Blumhagen R, Radune D, Hostetler JB, Brinkac LM, Durkin AS, Kluepfel DA, Wechter WP, Anderson AJ, Kim YC, 3rd Pierson LS, Pierson EA, Lindow SE, Kobayashi DY, Raaijmakers JM, Weller DM, Thomashow LS, Allen AE, Paulsen IT (2012) Comparative genomics of plant-associated *Pseudomonas* spp.: insights into diversity and inheritance of traits involved in multitrophic interactions. PLoS Genet 8(7):e1002784. doi:10.1371/journal.pgen.1002784, Epub 2012 Jul 5
Lopez-Velasco G, Welbaum GE, Boyer RR, Mane SP, Ponder MA (2011) Changes in spinach phylloepiphytic bacteria communities following minimal processing and refrigerated storage described using pyrosequencing of 16S rRNA amplicons. J Appl Microbiol 110:1203–1214
Lucero ME, Unc A, Cooke P, Dowd S, Sun S (2011) Endophyte microbiome diversity in micropropagated *Atriplex canescens* and *Atriplex torreyi* var griffithsii. PLoS One 6(3):e17693. doi:10.1371/journal.pone.0017693
Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S, Tremblay J, Engelbrektson A, Kunin V, del Rio TG, Edgar RC, Eickhorst T, Ley RE, Hugenholtz P, Tringe SG, Dangl JL (2012) Defining the core *Arabidopsis thaliana* root microbiome. Nature 488:86–90. doi:10.1038/nature11237
MacLean D, Jones JD, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Microbiol 7:287–296
Masson-Boivin C, Giraud E, Perret X, Batut J (2009) Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? Trends Microbiol 17:458–466
Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JH, Piceno YM, DeSantis TZ, Andersen GL, Bakker PA, Raaijmakers JM (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science 332:1097–1100

Monier JM, Lindow SE (2004) Frequency, size, and localization of bacterial aggregates on bean leaf surfaces. Appl Environ Microbiol 70:346–355

Nacke H, Thürmer A, Wollherr A, Will C, Hodac L, Herold N, Schöning I, Schrumpf M, Daniel R (2011) Pyrosequencing-based assessment of bacterial community structure along different management types in German forest and grassland soils. PLoS One 6(2):e17000

Nelson DM, Cann IKO, Mackie RI (2010) Response of archaeal communities in the rhizosphere of maize and soybean to elevated atmospheric CO_2 concentrations. PLoS One 5(12):e15897. doi:10.1371/journal.pone.0015897

Newton AC, Gravouil C, Fountaine JM (2010) Managing the ecology of foliar pathogens: ecological tolerance in crops. Ann Appl Biol 157:343–359

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA (1986) Microbial ecology and evolution—a ribosomal-RNA approach. Annu Rev Microbiol 40:337–365

Raaijmakers JM, Weller DM (1998) Natural plant protection by 2,4-diacetylphloroglucinol-producing *Pseudomonas* spp. in take-all decline soils. Mol Plant Microbe Interact 11:144–152

Rastogi G, Sbodio A, Tech JJ, Suslow TV, Coaker GL, Leveau JH (2012) Leaf microbiota in an agroecosystem: spatiotemporal variation in bacterial community composition on field-grown lettuce. ISME J 6:1812–1822

Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N (2010) The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. Environ Microbiol 12:2885–2893

Remus-Emsermann MN, Kim EB, Marco ML, Tecon R, Leveau JH (2013) Draft genome sequence of the phyllosphere model bacterium *Pantoea agglomerans* 299R. Genome Announc 1(1):e00036–13. doi:10.1128/genomeA. 00036-13

Rosenberg E, Sharon G, Atad I, Zilber-Rosenberg I (2010) The evolution of animals and plants via symbiosis with microorganisms. Environ Microbiol Rep 2:500–506

Sessitsch A, Hardoim P, Döring J, Weilharter A, Krause A, Woyke T, Mitter B, Hauberg-Lotte L, Friedrich F, Rahalkar M, Hurek T, Sarkar A, Bodrossy L, van Overbeek L, Brar D, van Elsas JD, Reinhold-Hurek B (2012) Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. Mol Plant Microbe Interact 25:28–36

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–1145

Socquet-Juglard D, Kamber T, Pothier JF, Christen D, Gessler C et al (2013) Comparative RNA-seq analysis of early-infected peach leaves by the invasive phytopathogen *Xanthomonas arboricola* pv. *pruni*. PLoS One 8(1):e54196. doi:10.1371/journal.pone.0054196

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A 103:12115–12120

Staley JT, Konopka A (1985) Measurements of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annu Rev Microbiol 39:321–346

Stassen JH, Seidl MF, Vergeer PW, Nijman IJ, Snel B, Cuppen E, Van den Ackerveken G (2012) Effector identification in the lettuce downy mildew *Bremia lactucae* by massively parallel transcriptome sequencing. Mol Plant Pathol 13(7):719–731. doi:10.1111/j.1364-3703.2011.00780.x, Epub 2012 Feb 1

Telias A, White JR, Pahl DM, Ottesen AR, Walsh CS (2011) Bacterial community diversity and variation in spray water sources and the tomato fruit surface. BMC Microbiol 11:81

Tikhonovich IA, Provorov NA (2011) Microbiology is the basis of sustainable agriculture: an opinion. Ann Appl Biol 159:155

Torsvik V, Øvreås L, Thingstad TF (2002) Prokaryotic diversity – magnitude, dynamics and controlling factors. Science 296:1064–1066

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC (2005) Comparative metagenomics of microbial communities. Science 308:554–557

Trivedi P, He Z, Van Nostrand JD, Albrigo G, Zhou J, Wang N (2012) Huanglongbing alters the structure and functional diversity of microbial communities associated with citrus rhizosphere. ISME J 6(2):363–383. doi:10.1038/ismej.2011.100, Epub 2011 Jul 28

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43

Uroz S, Buée M, Murat C, Frey-Klett P, Martin F (2010) Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. Environ Microbiol Rep 2:281–288

Vokou D, Vareli K, Zarali E, Karamanoli K, Constantinidou HI, Monokrousos N, Halley JM, Sainis I (2012) Exploring biodiversity in the bacterial community of the Mediterranean phyllosphere and its relationship with airborne bacteria. Microb Ecol 64:714–724

Wang HX, Geng ZL, Zeng Y, Shen YM (2008) Enriching plant microbiota for a metagenomic library construction. Environ Microbiol 10:2684–2691

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese C, Fox G, Zablen L, Uchida T (1975) Conservation of primary structure in 16S ribosomal RNA. Nature 254:83–86

Wu CH, Bernard SM, Andersen GL, Chen W (2009) Developing microbe–plant interactions for applications in plant-growth promotion and disease control, production of useful compounds, remediation and carbon sequestration. J Microbial Biotechnol 2:428–440

Yang CH, Crowley DE, Borneman J, Keen NT (2001) Microbial phyllosphere populations are more complex than previously realized. Proc Natl Acad Sci U S A 98:3889–3894

Yu X, Lund SP, Scott RA, Greenwald JW, Records AH, Nettleton D, Lindow SE, Gross DC, Beattie GA (2013) Transcriptional responses of *Pseudomonas syringae* to growth in epiphytic versus apoplastic leaf sites. Proc Natl Acad Sci U S A 29:E425–E434. doi:10.1073/pnas.1221892110

Zilber-Rosenberg I, Rosenberg E (2008) Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiol Rev 32:723–735

Chapter 10
Genes and *Trans*-Factors Underlying Embryogenic Transition in Plant Soma-Cells

Dhananjay K. Pandey and Bhupendra Chaudhary

In Vitro Regeneration of Plant Species

Reproduction is a unique feature of all living beings in order to maintain the species existence. In plant kingdom, majorly pollination, fertilization, embryo development and its germination are the natural practices to persist in the species management. Many plants choose alternate mechanisms for their genetic succession either parallel or independently. For example, in the genus *Bryophyllum* plantlets develop on the fringes of leaves whereas fragments of the plant bodies of liverworts and mosses regenerate through embryo formation. In many *Citrus* species, nucellar cells produce embryos, perhaps for overcoming the fertilization barrier and higher growth competency. Theoretically it is understood that all plant cells are totipotent and can give rise complete plantlet. If so, why do different taxa, cultivars and type of explants (tissues) exhibit significant variation in the regeneration potential? It is evident that different plant species respond *in vitro* with a wide spectrum of regeneration methods including organogenesis and somatic embryogenesis. The latter has drawn more attention in recent past as limited success had been achieved across plant species.

Somatic Embryogenesis

Somatic embryogenesis (SE) is one of the most diverged modes of regeneration in plants. Somatic embryogenesis was first reported in carrot cell suspension cultures (Steward et al. 1958). Somatic embryogenesis is the process of transition of a

D.K. Pandey, M.Tech. • B. Chaudhary, Ph.D. (✉)
School of Biotechnology, Gautam Buddha University,
Greater Noida 201310, Uttar Pradesh, India
e-mail: bhupendrach@gmail.com

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_10

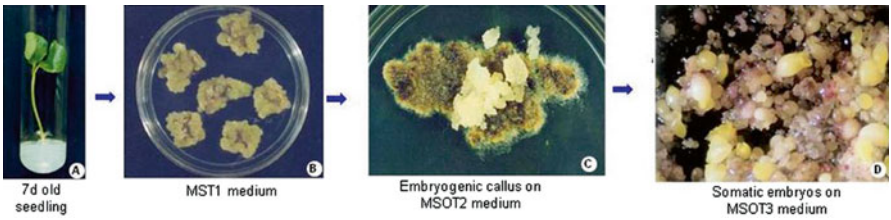

Fig. 10.1 Somatic embryogenesis in cotton (*Gossypium hirsutum* L. cv. Coker 310). (**A**) Cotyledonary explants were taken from 7 days seedling germinated on half-MS medium. (**B**) The explants produced calli on callus induction (MST1) medium. (**C**) The granular embryogenic callus on callus proliferation (MSOT2) medium. (**D**) Formation of embryogenic calli on embryo indcution (MSOT3) medium developing somatic embryos at this stage.

somatic cell into an embryo. In this process a bipolar structure is formed having a closed vascular supply making it independent from the parental tissue. This is a multi-step regeneration process in plants and is highly plant-genotype dependent. Previously, important roles of endogenous hormones have been suggested in such genotypic variations of SE potential (Carman 1990). Somatic embryogenesis shared many similarities with zygotic embryogenesis thus making it a model system for studying different biochemical, morphological and physiological characteristics in higher plants (Kawahara et al. 1995). Moreover, SE has wide use in genetic transformation, somatic hybridization and induction of somaclonal variations (Gall et al. 1994; Robertson et al. 1992). There are two modes of SE, one is direct somatic embryogenesis (DSE) characterized by the induction of somatic embryos directly from pro-embryogenic cells of leaves, stem, microspores or protoplasts without the proliferation of calli, and other is indirect somatic embryogenesis (ISE) where somatic embryos are developed from friable embryogenic calli (Jimenez and Bangerth 2001; Quiroz-Figueroa et al. 2006). In general, somatic embryo formation proceeds through distinct developmental stages involving globular, heart-shaped and torpedo-shaped stages in dicots (Fig. 10.1), and globular, scutellar (transition) and coleoptilar stages in monocots (Takano et al. 2006; Toonen et al. 1996).

It could thus be assumed that how much cellular changes have been required to bring such modulation in the cell-fate which was initially destined to multiply, but eventually converted into very specialized embryonic structures. What are those key factors which are responsible to compel a cell changing its fate? How could such factors be identified? Temporal study of the progression of SE at phenotypic and molecular level might be helpful in finding the precise and exact answers to such important questions.

Initiation of Somatic Embryogenesis

This is very crucial stage when the cellular machinery changes its programmed developmental paths. It is an induced cell-fate change triggered by different key factors in which applied stress is of prime importance. During evolution different

abiotic and biotic stress conditions are the prominent and determining factors of acquired cellular physiologies. Optimal stress conditions are required to trigger more cellular factors, which in turn are the part of different molecular mechanisms affecting the entire cellular milieu towards the acquisition of the embryogenic competence. For example, phytohormones 2, 4-D, ABA, ethylene are considered among such stress inducing factors. Gene expression study at transcriptome and proteome level has revealed the importance of selected stress related genes/proteins in the acquisition and maintenance of SE (Omid and Abbas 2010). Details of important stress-related factors for their roles towards SE acquisition have been described in upcoming sections.

Somatic Embryogenesis Is Genotype/Explant Source Dependent

Inducing a somatic cell to SE is highly dependent on the plant-genotype and the explants selected for *in vitro* regeneration (Maillot et al. 2006; Perrin et al. 2004). Many researchers reported that few cultivars had shown high regeneration potential in comparison with others of the same species (Rao et al. 2006; Trolinder and Goodin 1987). Similarly, within any particular regenerating cultivar differential magnitude of SE potential has been recorded when selected a typical tissue type as starting material (Perrin et al. 2004; Carimi et al. 2005; Gribaudo et al. 2004). In *Sorghum* genotype dependent SE and regeneration was reported when leaf base was considered as an explant (Mishra and Khurana 2003). Among various genotypes tested, CO 25 cultivar showed high frequency of callus induction and regeneration (Indra and Krishnaveni 2009). Genotype dependent SE in different varieties of rice has also been reported (Aruna and Reddy 1988; Moghaieb et al. 2009). Further, in sweet potato selection of specific genotype and explants has been observed to be important for induced ability of SE acquisition (Liu et al. 1992). In soybean among different genotypes tested on a wide range of media for SE, variation in frequency of embryogenesis induction was prominent (Bailey et al. 1993). Screening of 38 cultivars of the most important fibre crop cotton (*Gossypium*) for SE potential, suggested for very high degree of genotype specificity (Trolinder and Xhixian 1989). In Georgia and Pee Dee lines of cotton, regeneration efficiency is relatively low as compared to the standard Coker 312 cultivar and it shows a high degree of seed-to-seed variation in SE potential when tested *in vitro* (Sakhanokho et al. 2004). Moreover, in an elite Chinese cotton (*Gossypium hirsutum* L.) cultivar Simian-3, root is the most responsive explants for production of somatic embryos than hypocotyls and cotyledonary tissues (Zhang et al. 2001). It is important to note that the tissue types near to the zygotic embryo during developmental stage of a seed exhibited maximum response towards SE than that of distant tissues when used as explants. Probably, this is due to the existence of active cellular machinery responsible for induction and maintenance of embryonic potential in such tissue types. In sugarcane fresh inner whorl of leaves and shoot apical meristem has been found highly acquiescent to *in vitro* SE induction (Ali et al. 2007). Different peanut genotypes including Valencia, Spanish and Virginia botanical types showed significant

variations in somatic embryo formation and plantlet regeneration on similar culture conditions. Somatic embryo formation from cotyledons and embryo axis exhibited optimal embryogenic response supporting the fact of genotype and explant dependency during SE (Ozias-Akins et al. 1992).

Stress-Mediated Up-regulation of Phytohormone in SE

As noted elsewhere, plant SE is greatly dependent on varied stress conditions imposed *in vitro* across plant species (Feher et al. 2003; Kamada et al. 1989, 1993, 1994; Lee et al. 2001; Nolan et al. 2006; Touraev et al. 1997). Stress extensively interferes in activation of different cellular key factors during cell fate modulation. Until now, there are many reports highlighting important stress factors such as high osmotic pressure (Kamada et al. 1993; Iwai et al. 2003; Karami et al. 2006), explants wounding (Cheong et al. 2002; Santarem et al. 1997), dehydration (Kumria et al. 2003; Patnaik et al. 2005), heavy metal ions (Patnaik et al. 2005; Cueva Agila et al. 2013), high temperature (Kamada et al. 1994; Kiyosue et al. 1993a), micro- and macronutrients (Pandey et al. 2012) eliciting SE acquisition response in differential manner. However, the major question remains whether different stress conditions trigger identical cellular pathways or differentially contribute in SE acquisition. Therefore, it would be remarkable to explore the molecular mechanisms of such stress mediated SE acquisition. Work done so far in this area suggested that reprogramming of cellular fate is due to pattern change in the expression of candidate genic- factors in response to diverse stress conditions. Such important factors may in turn trigger the cellular modulation for embryo formation (Feher et al. 2003; Chugh and Khurana 2002; Ge et al. 2012; Namasivayam 2007; Rose and Nolan 2006; Schlögl et al. 2012; Wiśniewskaa et al. 2012; Wójcikowska et al. 2013; Xu et al. 2013; Zheng et al. 2013). It is evident that phytohormones' level is critical for the induction of SE under stress conditions, which subsequently may influence genetic and cellular *trans*-factors essential for cell-restructuring and re-programming. Essential phytohormones studied so far for their definite role in the induction of SE are as following:

Auxin

In most of studied plant taxa, auxin has been reported to be the prime factor for somatic cell to embryo transition (Jimenez and Bangerth 2001; Ivanova et al. 1994; Michalczuk et al. 1992; Rajasekaran et al. 1987). An optimal amount of auxin is necessary for the induction and progression of SE. At initial stage of SE high amount of auxin is required that may be exogenously supplied *in vitro* cultures. However, at later stages of embryo formation, endogenous auxin-level is declined gradually establishing an auxin gradient in the soma-cell considered to be important for

auxin-mediated cell-signalling (Michalczuk et al. 1992; Fujimura and Komaminea 1979; Gray et al. 1984; Gray and Trigiano 1993; Hanning and Conger 1982). In parallel with zygotic embryos it is evident that the polar transport of auxin is indispensable for the establishment of bilateral symmetry in dicotyledonous somatic embryos (Liu et al. 1993; Schiavone and Cooke 1985, 1987). The endogenous level as well as exogenously supplied auxin acts as one of the determining factors during induction and maintenance of SE (Michalczuk et al. 1992). Therefore, the threshold level of auxin has to be maintained inside the competent somatic cell for the acquisition of embryogenic potential. Significant role of auxin has been proved in *Pennisetum purpureum* (Rajasekaran et al. 1987), sugarcane (Guiderdoni et al. 1995), carrot (Jiménez and Bangerth 2001a; Li and Neumann 1985; Sasaki et al. 1994), *Medicago falcata* (Ivanova et al. 1994), *Prunus* spp. (Michalczuk and Druart 1999), maize (Jiménez and Bangerth 2001) and wheat (Jiménez and Bangerth 2001b) measuring higher endogenous auxin level in the embryogenic competence (EC) in comparison with non-embryogenic competence (NEC) cell. Further it has been suggested that different stress response is directly linked with auxin and auxin response factors. Recently, important role of micronutrient mediated stress in maintenance of optimal auxin level in cellular milieu leading to SE acquisition has been suggested (Pandey et al. 2012). Differential expression of auxin-responsive genes established a relationship between stress and downstream gene expression (Jain and Khurana 2009). However, it becomes essential to reveal the molecular mechanism of stress-mediated up-regulation of auxin-responsive gene network with their functional components. Expression analyses have established relationship between different type of stress conditions and auxin with downstream factors essential to the acquisition of cell-to-embryo transition. For example, enhanced endogenous auxin level upregulates the expression of somatic embryo receptor kinase (*SERK*) gene (Nolan et al. 2006), somatic embryo-related factor (*SERF*) (Mantiri et al. 2008) and Ca^{2+} ion channel-mediated regulatory gene expression (Takeda et al. 2003), all reported earlier to be essential for SE. Auxin-inducible genes in *Triticum aestivum* including *Ta*SERKs are involved in SE and decreased significantly upon auxin depletion (Singla et al. 2008). Under various stress conditions in rice, there is relation between endogenous IAA levels and expression of genes related to biosynthesis/signalling of the hormone (Du et al. 2013). These studies confirmed the strong association among stress, auxin and triggered key factors considered to be essential for somatic cell-to-embryo transition. The auxin response factors (ARFs) are the imperative determining factors which play significant role in the downstream effect of auxin in the cellular milieu. So far many ARFs have been discovered for their vital role in SE, for example ARF7 and ARF19 proteins regulate target genes of auxin cell-signaling during embryonic development. These observations provide molecular insight into the unique contribution of auxin towards plant SE (Okushima et al. 2005; Wang et al. 2007). Further, the next generation sequencing (NGS) technology might help in sequencing of ARFs and their downstream target-expression profiling in the maintenance of SE (Yang et al. 2012).

Cytokinin

Cytokinin acts as an important growth regulator for somatic embryos commencement *in vitro* and has already been reported in many species such as *Trifolium*, *Coffea*, *Gladiolus*, *Helianthus*, *Spinacia* and *Medicago*. The immature zygotic embryos of *Trifolium* when cultured in presence of phytohormone BAP, cells of hypocotyl epidermis directly produce somatic embryos without an intermediate callus stage (Maheswaran and Williams 1985). Leaf explants of *Coffea* when cultured on high cytokinin containing medium, white friable calluses were produced and consequently developed into somatic embryos. Surprisingly, subculturing and maintenance of calli on the same medium produced somatic embryos even after 2 years of sub-culturing (Yasuda et al. 1985). In high cytokinin containing suspension cultures of *Gldiolus* primary and secondary embryos could be developed from a single somatic cell (Remotti 1995). Zeatin, a cytokinin also had a positive effect on embryogenic potency of *Helianthus* during *in vitro* culture (Charrière and Hahne 1998). In wild *Medicago* spp. (*M. truncatula*, *M. littoralis*, *M. murex*, *M. polymorpha*) high cytokinin content significantly supports the progression of SE (Iantcheva et al. 1999).

Abscisic Acid (ABA)

In plants, many physiological processes are mediated and regulated by ABA and has also been reported to be important for embryogenesis (Dodeman et al. 1997; Kong and Von Aderkas 2007; Vahdati et al. 2008). Stimulatory consequence of ABA on somatic cells or tissues modulates its fate towards embryogenesis and is considered as a stress response (Zdravković-Korać and Nešković 1999). ABA increases desiccation tolerance of the tissue destined to embryo formation and in the process increases the competence of embryo formed. Expression of specific genes and proteins, for example LEA, *Dc*ECP31 (Kiyosue et al. 1992), *Dc*ECP40 (Kiyosue et al. 1993b) in carrot, *At*ECP31 (Yang et al. 1996), *At*ECP63 (Yang et al. 1997) in *Arabidopsis*, *Leu* D19 in cotton and *Em* gene in wheat (Dure et al. 1989; Hughes and Galau 1991; Marcotte et al. 1988) were examined up-regulated during SE and identified as a component of ABA-inducible system. In support to this, ABA supplied exogenously to the non-embryogenic callus showed stimulation for the SE (Cui et al. 1998). ABA may also regulate the embryogenesis process mainly through regulation of *DC*8 genes (Hatzopoulos et al. 1990), carbohydrate metabolism (Karami et al. 2006; Martin et al. 2000), inhibition of peroxidase activity (Wei 2001) and others. These studies highlight the predominant role of ABA in the process of SE.

Gibberellins

The role of exogenous and endogenous gibberellins in SE has limited intervention so far. Preliminary results highlighted the high levels of endogenous gibberellins in the initial explants and the callus cultures are considered to be important for both

embryo induction and elongation (Jiménez 2001). Moreover, effects of gibberellins on SE depends on the genotype and explants used for tissue culture (Mikuła et al. 1998). For example, exogenous application of a subtype of gibberellin GA3 in culture medium inhibited SE in carrot (Fujimura and Komamine 1975; Tokuji and Kuriyama 2003), citrus (Kochba et al. 1978) and *Geranium* (Hutchinson et al. 1997). Whereas exogenous application of GA3 stimulated embryogenesis in immature cotyledon cultures of *Cicer* (Hita et al. 1997) and petiole derived cultures of *Medicago* (Ruduś et al. 2002).

Ethylene

Endogenous ethylene has a vital role in the conversion of soma-cells into somatic embryos, thus fine balance of ethylene level is essential for optimal embryogenic competency (Tsuchisaka et al. 2009). In the near iso-genic lines (NILs) of white spruce for SE, the ethylene content was observed relatively higher in the non-embryogenic line (Kumar et al. 1989) and exogenous supply of the same led to the formation of abnormal embryo (Kong et al. 1999; Kong and Yeung 1994). Recent studies confirmed that high ethylene levels at maturation stage of SE inhibit somatic embryo development and, reduced ethylene-level further enhanced embryo formation (Huang et al. 2001; Minocha et al. 2004; Ptak et al. 2010). Similar to gibberellins, response of ethylene towards SE is highly genotypic dependent and varied across species and tissue types (Jiménez 2005). For example, addition of ethylene inhibitors enhanced embryogenic-callus initiation in *Zea mays* (Vain et al. 1989) and SE in *Picea glauca* (Kong and Yeung 1994), *Hevea brasiliensis* (Auboiron et al. 1990). On contrary, SE has been shown to be stimulated by ethylene in *Medicago sativa* (Kepczynski et al. 1992) and *Coffea camephora* (Hatanaka et al. 1995). Moreover, in *Daucus carota* ethylene may either act as an inhibitor (Roustan et al. 1990) or stimulator (Nissen 1994). Thus the effect of ethylene on the induction of SE varied within and across species.

Induced Cell-Fate For SE

Cellular Morphology, Physiology and Histological Pattern

During SE, somatic cells convert into embryonic tissue and are of wide importance for the study of morphological changes occurred during the process. Changes arose during SE induction provide clues to understand the mechanism of transition from an un-differentiated to embryogenic tissue. Following developmental stages, embryonic competence is achieved by somatic cells similar to zygotic cells however, with certain variations at cellular and histological levels.

In many plant species, cells involved in SE are small in size and have large nucleus, thick cytoplasm containing numerous ribosomes, mitochondria, plastids with starch and short profiles of rough endoplasmic reticulum (Canhoto et al. 1996;

Zhang et al. 1997). In Pineapple Guava (*Myrtaceae*), sub-epidermal cells of the upper cotyledonary surface undergo several divisions giving rise to multi-meristematic layers which subsequently developed into somatic embryos (Canhoto et al. 1996). Also, plasmodesmata and oil bodies during early developmental stages of embryos considered to be important for proper exchange of material and nutritional supply required for embryo maturation (Canhoto et al. 1996). Epithelial cells were observed important for SE induction in rice scutellum explants, since it acts as main player in absorption of sugar, phytohormones and having high metabolic activity (Jones and Rost 1989; Maeda and Radi 1991). The microtubule organization of the cell wall is also important during SE mainly through the maintenance of cell size (Toonen et al. 1996). Also the culture density of cell suspensions could be a significant factor as low culture density has been shown favouring SE (Fujimura and Komaminea 1979; Nomura and Komamine 1985). Somatic embryogenesis is extensively inhibited in high cell density cultures of carrot cells, and used to identify the inhibitory factor 4-hydroxybenzyl alcohol that could strongly inhibit the formation of somatic embryos when added to the culture medium (Kobayashi et al. 2000). Contents of reducing sugars and starch have also been reported as one of the important factors in differentiation of embryogenic and non-embryogenic calli. For example, in *Medicago arborea*, higher sugar concentrations and lower starch content in the embryogenic cultures are prevalent than in non-embryogenic cultures (Martin et al. 2000). Various studies confirmed that carbohydrate in form of starch has high consumption in SE competent cells and therefore, may be considered as a marker for SE induction (Ho and Vasil 1983; Radojevic 1979). Furthermore, callose deposition in the cell wall and the presence of vacuolar Ca^{2+} may also be considered as the primary signals for the identification of embryogenic competent cells, as shown in *Cichorium* (Dubois et al. 1991) and *Trifolium repens* (Maheswaran and Williams 1985). From the cell suspensions of carrot taking an individual cell and the monoclonal antibody JIM8 for a cell wall antigen (McCabe et al. 1997) have shown that $JIM8^{+}$ cell population produced somatic embryos, whereas $JIM8^{-}$ population could not develop any embryos. Physical isolation of competent cells is also considered to be important for SE induction, as it helps in perceiving altered signalling and behavioural independency from neighbouring cells. The concept of such isolation holds strong from the finding that competent cells subsequently loss their plasmodesmata between surrounding cells, interrupting symplastic continuity and reducing electrical coupling (Warren and Warren 1993).

In *G. hirsutum* the histological examination showed epidermal or sub-epidermal types of cells get induced during embryogenesis. The embryogenic mass and somatic embryos are mostly derived from these sub-epidermal cells only. The chromosomal behaviour of the cells surrounding embryogenic cells often showed abnormal divisions and higher level of polyploidy (Aydin et al. 2010). In Pineapple Guava (*Myrtaceae*) 4–5 day culture of cotyledons when exposed to SE induction medium, the accumulation of starch in the plastids and the differentiation of mitochondria is the most pronounced modifications (Canhoto et al. 1996). During maturation, in many cases, the embryos undergo structural development and also accumulate proteinaceous and carbohydrate compounds necessary for further development (Brownfield et al. 2007; Tereso et al. 2007).

Changes in Gene Expression

Somatic cells are guided by a set of genes and proteins having their precise temporal expression profiles. The molecular switch of a somatic cell into embryo is the diversion of the existing genetic program. During transition of somatic cell into embryo, the expression pattern of such candidate genes changes so that cell under conversion may attain an alternative physiology. Such reprogramming could occur at transcriptional or post-transcriptional levels. It has been earlier reported that SE requires increased RNA synthesis during the molecular switch (Fujimura and Komamine 1980). Also exogenous and endogenous levels of corresponding proteins in culture directly affect the induction of SE (de Vries et al. 1988; Gavish et al. 1991). Therefore, identification of candidate genes will help to understand the mechanism of somatic cell-to-embryo transition at cellular level. Previously candidate genes have been examined to express exclusively during the embryogenic stage, and such genes could easily be used as robust biomarkers of SE.

Somatic Embryo Receptor Kinase (*SERK*)

Somatic embryo receptor kinase (*SERK*) is highly potential candidate gene and has been shown previously to have direct relation with SE. Over-expression of *SERK* gene exhibited three- to fourfold increase in the embryogenic competence across plant taxa (Hecht et al. 2001). At first, SERK were identified as *trans*-membrane protein at the cell surface of *D. carota* cell cultures (Schmidt et al. 1997). It has been observed that these molecules are highly expressed in vacuolated cells. Up to globular stage of embryogenesis, *SERK* gene expression has been examined high in both somatic and zygotic embryos. Until now there are many reports from different taxa supporting the fact that *SERK* gene expression is important for SE. *SERK* genes positively regulate SE in rice (Hu et al. 2005), *Arabidopsis* (Hecht et al. 2001), potato (Sharma et al. 2008), cotton (Pandey and Chaudhary 2014);coconut (Perez-Nunez et al. 2009), *Medicago* (Nolan et al. 2003), orchid species *Cyrtochilum loxense* (Cueva et al. 2012), sunflower (Thomas et al. 2004), cacao (Santos et al. 2005), grapevine (Maillot et al. 2009) and gymnosperm *Araucaria* (Steiner et al. 2012). Homologs of *SERK* gene have been discovered in several plant species with their role in SE(Baudino et al. 2001; Ito et al. 2005; Shimada et al. 2005). In plant, the SERK over-expression has been shown to activate several molecular signals/ligands at the cell-surface. Possibly, the molecular signals/ligands develop a LRR-mediated attachment to SERK inducing the directed signaling cascade. In result, the concurrent expression network get modulated possibly via chromatin remodeling further inducing other target genes responsible for SE, e.g. LEA, LEC, BBM and their combinatorial effect compel a somatic cell to get converted into an embryo (Chugh and Khurana 2002; Schmidt et al. 1997; Maillot et al. 2009; Albrecht et al. 2008; Braybrook et al. 2006; Casson et al. 2005; Heidmann et al. 2006; Ikeda et al. 2006; Nolan et al. 2009; Passarinho et al. 2008; Zhenga et al. 2009). In cotton micronutrient boron mediated *in vitro* stress condition enhanced endogenous auxin level in the

somatic cells, resulting into chromatin-remodeling and up-regulation of *SERK* transcript level (Albrecht et al. 2008). All these findings conclude that *SERK* could be considered a molecular marker of SE in plants.

WUSCHEL (*WUS*)

WUSCHEL is an important gene known for its specific role in the determination of cellular fate in plant regeneration mainly through SE (Namasivayam 2007). Previous reports largly emphasized the role of *WUS* gene during *in vitro* regeneration (Atta et al. 2009; Cary et al. 2002; Gordon et al. 2007). *WUS* gene has been examined for the promotion of vegetative-to-embryonic transition in *Arabidopsis* (Zuo et al. 2002). Optimum expression of *WUS* gene regulated by an appropriate amount of exogenously supplied auxin is indispensable for acquisition of SE (Zuo et al. 2002). It has also been suggested that the establishment of auxin gradient is fairly correlated with the induced *WUS* expression during SE (Su et al. 2009). Recently, it has been shown that over-expression of the *WUS* gene from *Arabidopsis* enhanced embryogenic competence in cotton callus culture *in vitro* (Bouchabké-Coussa et al. 2013). In *Ocotea catharinensis*, an endangered angiosperm tree species, *WUS-related* genes have been observed up-regulated during embryogenesis (Santa-Catarina et al. 2012). *WUS* expression is primarily regulated by feedback loop involving *CLV* genes (Bhalla and Singh 2006), and mutation in *CLV* genes leads to over-expression of *WUS* due to lack of feedback control. Therefore, it is important to investigate such factors having suppressive effect on *CLV* genes those might possibly be important for SE directly or indirectly through *WUS* over-expression. So far, in many plant species *WUS* gene expression has been observed important for the induction of SE.

Baby Boom (*BBM*) Gene

The *BBM* gene was first isolated from microspore cultures of *Brassica napus* (Boutilier et al. 2002). BBM is a transcription factor of AP2/ERF family expressed preferentially in seed, embryo and root meristem (Nole-Wilson et al. 2005). In *Arabidopsis* and *Brassica*, ectopic expression of *BBM* triggers a switch from vegetative to embryonic growth (Boutilier et al. 2002; Kulinska-Lukaszek et al. 2012). Heterologous expression of BABY BOOM AP2/ERF transcription factor from *B. napus* and *A. thaliana* under constitutive CaMV 35S promoter enhanced regeneration potential of tobacco (Srinivasan et al. 2007). This suggests that *BBM* is an important player during acquisition of plant SE. However, it is of prime concern what are the downstream signalling components of BBM-induced SE? Using yeast-two-hybrid techniques certain BBM interacting proteins have already been identified, for example PKR1 (PICKLE-RELATED1) and HDG11 (HOMEODOMAIN

GLABROUS11). Though, further experimental validation shall be required to study the loss or gain-of-function mutants of PKR1 and/or HDG11, to characterize the role of *BBM* in somatic embryos formation (Horstman et al. 2009).

WRKY, *AOX* and Ca^{2+}

WRKY gene family is one of the largest families of transcriptional regulators in plants and participate in signalling network leading to wide and differential functions. Ongoing research in this area suggested that the family had evolved with members of diverse functionality showing propoerties of both activators and repressors of target genes performing different cellular processes. Moreover, it is interesting to note that single WRKY transcription factor may act in differential manner in varied cellular process and conditions. During SE the WRKY transcription factor DGE1 in orchard grass (*Dactylis glomerata*) has been examined over expressed (Alexandrova and Conger 2002). In *Solanum chacoense* the WRKY gene family member(*Sc*WRKY1) consisted of 525 amino acids and play specific role during embryogenesis (Lagace and Matton 2004). Also, in response to different stress conditions increased expression level of *WRKY* gene was recorded and further considered to be important in the regulation of transcriptional reprogramming required for SE (Lagace and Matton 2004; Chen et al. 2011; Kasajima et al. 2010; Zou et al. 2004). Though the initial data suggested for the essential role of WRKY in SE, target family members have yet to be experimentally validated for their precise function(s) through specific molecular interactions. Under stress conditions, the somatic cells accumulate large number of oxidant molecules, for example, micronutrient Boron (B)-deficiency in cultured somatic cells plasma membranes were highly leaky and lose their functional integrity. Under such conditions, the soma-cells also had impaired plasma membranes with disturbed ion fluxes, decreased level of photosynthesis and poor defense system (Cakmak and Römheld 1997). However, to sustain the cell-vigour up-regulation of mitochondrial alternative oxidase (*AOX*) gene was observed enhancing the defense mechanism of the cell mainly through lowering down the reactive oxygen production (Maxwell et al. 1999). Similarly, in carrot two *AOX* genes (*Dc*AOX1a and *Dc*AOX2a) have been examined up-regulated during initiation of SE (Frederico et al. 2009).Thus the *AOX* gene family has been identified having crucial role in plant SE.

Ca^{2+} is one of the most important signalling molecules participating in many cellular pathways and transmitting different signals. The location, concentration and persistence time of Ca^{2+} mediated signals are the determining factor for its signalling diversity. Intracellular Ca^{2+} concentration has been observed up-regulated during embryogenesis, indicating its active role in this process (Anil and Rao 2000; Antoine et al. 2000; Jansen et al. 1990). During SE Ca^{2+} ion dependent protein kinase (CDPKs) acts as key player in the signalling pathway as the CDPKs and significant levels have been traced in the embryogenic cell culture's protein extracts (Jackson and Casanova 2000; Steenhoudt and Vanderleyden 2000).

Altered Cellular Homeostasis Is Essential for Soma Cell-to-Embryo Transition

A pre-requisite to the induction of embryogenic potential is the treatment of soma cells under induced stress conditions subsequently up-regulating cell-signalling (Fig. 10.2). However, prolonged treatment of soma-cells under induced stress conditions may result into the damage of cellular machinery and decreased SE. The Reactive Oxygen Species (ROS) are mostly the oxidants molecules released in the cell during induced stress conditions or developmental processes. In stress conditions, these may negatively influence the soma cell-to-embryo conversion process. In order to avoid the negative impact of such oxidant molecules, the cells have been evolving with the mechanisms of scavenging excess ROS produced in the growing cell preventing cellular damage. This may involve one of the important processes of up-regulation of antioxidant defense mechanisms (Apel and Hirt 2004; Bowler et al. 1992; Willekens et al. 1997). Among all antioxidant molecules, glutathione is the primary molecule identified to prevent the ROS-mediated cellular damage thus maintaining the cellular homeostasis (Noctor and Foyer 1998). Previously, it has been proved that the certain enzymes targeting glutathione degradation in the cell are up-regulated during auxin-induced SE. If so, up-regulation of antioxidant glutathione may help in achieving the homeostasis much earlier than the required period for the acquisition of SE. Earlier, it has been shown that glutathione degrading enzymes are prominently expressed during auxin-induced SE in *Cichorium* in order to maintain the auxin-mediated induction of embryonic siganling cascade and thus delaying of achieving the cellular homeostasis (Galland et al. 2001; Thibaud-Nissen et al. 2003). However, under induced stress conditions an auxin gradient is established in the cell by the coordinated expression of these enzymes (Nagata et al. 1994). It is thus concluded that embryogenic cells do control oxidative stress much efficiently by regulating the ROS-scavenging system of the cell. Therefore the soma cells having stress-induced signaling cascade upto a threshold level shall only be competent to cell-to-embryo transition. But in response to the increased oxidants in the cell, the antioxidant level will also be increased maintaining the cellular redox (Fig. 10.2) (Kairong et al. 1999). It is thus evident that induction of SE is a result of modulated ROS levels majorly through balancing of stress-mediated production of oxidants and in turn up-regulated antioxidants (Pandey and Chaudhary 2014). Hence, it may be assumed that in order to maintain cellular homeostasis, even the optimal levels of ROS essential in SE will be scavenged. Since the competent somatic cells entail stress-mediated optimal ROS levels for the induction of SE (Fig. 10.2), removal of ROS may also elude the embryogenic potential *in vitro*.

Genomics of Somatic Embryogenesis

The advent of genomics with newly emerged technologies such as Next Generation Sequencing (NGS) platforms commenced a new dimension of studying genetic information of complex biological system. It is now easy to study the genome,

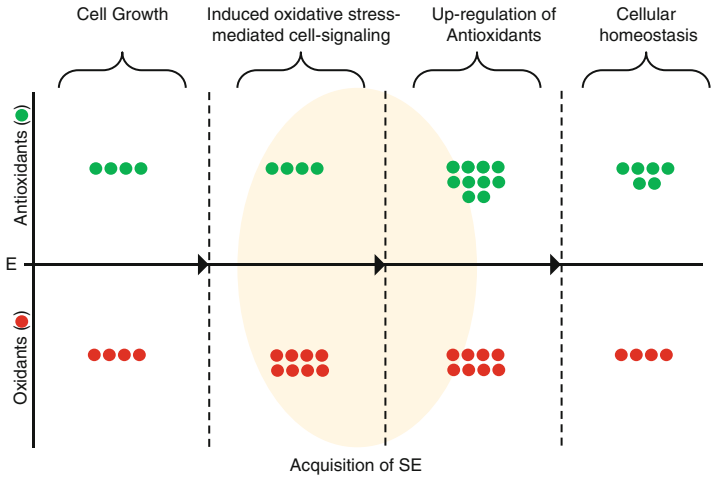

Fig. 10.2 A hypothetical model revealing the alteration and maintenance of cellular redox during the induction of SE. During cellular development a fine balance of oxidant (red circles) and antioxidant (green circles) molecules is maintained by the soma-cells. Under induced stress conditions, the stress-mediated cell-signalling is up-regulated and is the pre-requisite for the acquisition of SE. During this period (as highlighted in grey color) most of the soma-cells exposed to stress conditions undergo molecular and cellular changes leading to the formation of somatic embryos. However, the prolonged exposure to applied stress conditions, cellular defense mechanism mainly through the up-regulation of antioxidant genes is up-regulated and leads to the maintenance of cellular homeostasis. At this stage, the threshold levels of cell-signalling molecules are suppressed and the acquired embryogenic potential is lost.

transcriptome and epigenome of any species with wider approach and provides new information to understand the complex mechanism of diversity of cellular function in the biological systems. NGS platforms may also be used as powerful tool to study the molecular basis of SE. Using NGS technologies now it is easy to compare variations in the transcriptome of embryogenic and non-embryogenic stages/cultivars of a plant species. For example, in cotton two near-isogenic lines for the trait of full-regeneration (FR) and non-regeneration (NR) could be compared at transcript level to study the molecular basis of acquired SE.

Limited research information is available in the area of plant SE using transcriptome sequencing platforms. Earlier, in cotton, 137 differentially expressed genes were either switched on or off during transition of non-embryogenic calli to embryogenic calli, whereas 813 genes changed their expression level during the development process of somatic embryos (Yang et al. 2012). *In silico* analysis shown that the genes for basic processes such as metabolic pathways and biosynthesis of secondary metabolites are up-regulated in embryogenic calli than non-embryogenic calli. Cell wall related proteins, for example arabinogalactan protein and aquaglyceroporin along with diverse range stress have already been reported earlier for their importance in SE acquisition. Microarray studies had shown up-regulation of auxin and related gene network which included indole-3-acetate beta-glucosyltransferases, nitrilase of IAA metabolism and IAA18, auxin-responsive GH3-like protein for

auxin regulation pathway (Seki et al. 2002). Variation in the transcript levels of genes was observed related to auxin biosynthesis, conjugate metabolism and transportation (at least 35 transcripts), auxin-response factors & responsive elements (at least 36 trasncripts) and other auxin-related proteins (11 transcripts) during SE. Moreover, complex expression patterns throughout SE in cotton possessed tryptophan biosynthesis 1 (TRP1),tryptophan synthase β-subunit 2 (TSB2), chorismate mutase (CM1), CYP79C1, YUC and FMO, IAA biosynthesis transcripts, and nitrilase 4 (NIT4). *TRP1* and *ASB1* have been observed up-regulated throughout embryogenesis, while *NIT4A* was down-regulated, confirming the role of auxin in SE (Yang et al. 2012). Similarly, more than 400 diverse transcription factors' (TFs) mRNAs expressed during SE. During late embryogenesis, number of TF mRNAs decreased over time. Interestingly, Zinc finger and bHLH family TFs are highly expressed among all studied TFs. The MYB family TFs and others such as ERF, bZIP and WRKY accounted for 4–6 % of identified transcripts supporting for their direct contribution in induced SE. Most of the unigenes identified are linked to the functions of protein binding, hydrolase activity, phytohormone signalling, cell wall and cell membrane modification and abiotic/biotic stress (Yang et al. 2012). Recently, conserved and species-specific microRNAs have also been identified over-expressed during SE in Poplar using deep sequencing and microarray hybridization (Li et al. 2012). The results suggested about potential miRNA targets and their functions which is directly linked to diverse biological and metabolic processes. Their predicted target genes are mainly involved in many metabolic and biological processes including signal transduction, protein metabolism, responses to abiotic or biotic stimulus, growth, cell organization, electron transport and energy pathways, and many other developmental processes. Further, the molecular functions of these target genes included DNA or RNA binding, nucleotide binding, involvement in enzyme activity, receptor binding, ATP binding and others. However, the function of most of the miRNA target genes are still unknown(Li et al. 2012).

In cotton, small RNA and degradome sequencing confirmed complex miRNA regulation during SE. Thirty-six known miRNA families were examined to be differentially expressed and 25 novel miRNAs were identified involved in this process (Yang et al. 2013). In the two lines of cotton, global analysis of transcriptome dynamics was performed using RNA-Seq during SE and a total of 204,349 unigenes were identified by de novo assembly of the 214,977,462 Illumina reads (Xu et al. 2013). The expression of phytohormone-related genes was mainly linked with auxin and cytokinin biosynthesis and signal transduction pathways. Because the concentration ratio of phytohormone auxin and cytokinin supplied in culture media and their subsequent increase in endogenous levels at different developmental stages of regeneration process directed the induction of SE (Xu et al. 2013). During the global transcriptome analysis of near isogenic lines of plant species for the trait of SE, expression status of key *responsible genes/factors* and their master regulators will provide magnificent outlook on the molecular mechanism behind SE. Validation of already reported genes/transcription factors and exact consequences of their expression during SE acquisition would be feasible. Finding out the effect on diverse gene network during different stages of SE acquisition and maintenance would help in

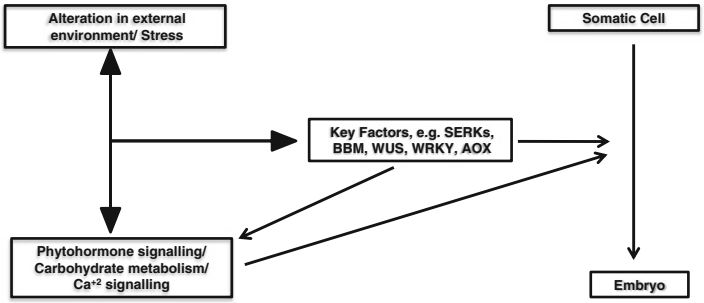

Fig. 10.3 Diagrammatic representation of possible interactions at cellular and molecular levels during somatic cell-to-embryo transition process. Modern high throughput technologies have supported the analyses of intricate mechanisms responsible for embryonic phenotype across species. The application of genomics tools such as NGS platforms would be immensely beneficial in the elucidation of the most consensus mechanism that may link temporal cellular physiologies during the acquisition and maintenance of somatic embryogenesis.

establishing direct and/or indirect correlation between diversity of cellular factors/processes and its contribution towards SE (Fig. 10.3). Correlation between varied stress factors and culture condition and its impact on transcript level leading a cell towards SE is now easy to measure using different NGS platforms. Efforts have been made to categorize the transcript expression for phytohormone auxin and cytokinin in two distinct lines of cotton and differential expression levels were recorded along with the expression variation of other genes (Xu et al. 2013). There are more important factors and molecular mechanisms participate in this complex process of cellular fate-change yet to be examined which might help in our understanding of the complex network of induced SE. Certainly NGS technologies is the new hope in the scientific community working in the area of plant regeneration biology.

References

Albrecht C, Russinova E, Kemmerling B, Kwaaitaal M, de Vries S (2008) *Arabidopsis* somatic embryogenesis receptor kinase protein serves brassinosteroid-dependent and independent signalling pathway. Plant Physiol 148:611–619

Alexandrova KS, Conger BV (2002) Isolation of two somatic embryogenesis-related genes from orchard grass (*Dactylis glomerata*). Plant Sci 162:301–307

Ali A, Naz S, Iqbal J (2007) Effect of different explants and media compositions for efficient somatic embryogenesis in sugarcane (*Saccaharum officinarum*). Pak J Bot 39:1961–1977

Anil V, Rao K (2000) Calcium-mediated signaling during sandalwood somatic embryogenesis. Role for exogenous calcium as second messenger. Plant Physiol 123:1301–1311

Antoine A, Faure J, Cordeiro S, Dumas C, Rougier M, Feijo J (2000) A calcium influx is triggered and propagates in the zygote as a wavefront during in vitro fertilization of flowering plants. Proc Natl Acad Sci U S A 97:10643–10648

Apel K, Hirt H (2004) Reactive oxygen species: metabolism, oxidative stress, and signal transduction. Annu Rev Plant Biol 55(1):373–399

Aruna M, Reddy GM (1988) Genotypic differences in callus initiation and plant regeneration from anthers of indica rice. Curr Sci 57:1014–1017

Atta R, Laurens L, Boucheron-Dubuisson E, Guivarc'h A, Carnero E, Giraudat-Pautot V et al (2009) Pluripotency of Arabidopsis xylem pericycle underlies shoot regeneration from root and hypocotyl explants grown in vitro. Plant J 57:626–644

Auboiron E, Carron M-P, Michaux-Ferrière N (1990) Influence of atmospheric gases, particularly ethylene, on somatic embryogenesis of *Hevea brasiliensis*. Plant Cell Tiss Org Cult 21:31–37

Aydin Y, Talas-Oğraş T, Altinkut A, Ismailoğlu I, Arican E, Gozukirmizi N (2010) Cytohistological studies during cotton somatic embryogenesis with brassinosteroid application. IUFS J Biol 69:33–39

Bailey MA, Boerma HR, Parrott WA (1993) Genotype effects on proliferative embryogenesis and plant regeneration of soybean. In Vitro Cell Dev Biol – Plant 29:102–108

Baudino S, Brettschneider R, Hecht V, Dresselhaus T, Lorz H, Dumas C et al (2001) Molecular characterization of novel maize LRR receptor-like kinases, which belong to the SERK family. Planta 213:1–10

Bhalla PL, Singh MB (2006) Molecular control of stem cell maintenance in shoot apical meristem. Plant Cell Rep 25:249–256

Bouchabké-Coussa O, Obellianne M, Linderme D, Montes E, Maia-Grondard A, Vilaine F et al (2013) Wuschel overexpression promotes somatic embryogenesis and induces organogenesis in cotton (*Gossypium hirsutum* L.) tissues cultured in vitro. Plant Cell Rep. doi:10.1007/s00299-013-1402-9

Boutilier K, Offringa R, Sharma VK, Kieft H, Ouellet T, Zhang LM et al (2002) Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. Plant Cell 14:1737–1749

Bowler C, Montagu MV, Inze D (1992) Superoxide dismutase and stress tolerance. Annu Rev Plant Physiol Plant Mol Biol 43(1):83–116

Braybrook S, Stone S, Park S, Bui A, Le B, Fischer R et al (2006) Genes directly regulated by LEAFY COTYLEDON2 provide insight into the control of embryo maturation and somatic embryogenesis. Proc Natl Acad Sci U S A 103:3468–3473

Brownfield L, Ford K, Doblin MS, Newbigin E, Read S, Bacic A (2007) Proteomic and biochemical evidence links the callose synthase in *Nicotiana alata* pollen tubes to the product of the NaGSL1 gene. Plant J 52:147–156

Cakmak I, Römheld V (1997) Boron deficiency-induced impairments of cellular functions in plants. Plant Soil 193:71–83

Canhoto JM, Mesquita JF, Cruz GS (1996) Ultrastructural changes in cotyledons of pineapple guava (Myrtaceae) during somatic embryogenesis. Ann Bot 78:513–521

Carimi F, Barizza E, Gardiman M, Lo Schiavo F (2005) Somatic embryogenesis from stigmas and styles of grapevine. In Vitro Cell Dev Biol – Plant 41:249–252

Carman JG (1990) Embryogenic cells in plant tissue cultures: occurrence and behavior. In Vitro Cell Dev Biol 26:746–753

Cary A, Che P, Howell S (2002) Developmental events and shoot apical meristem gene expression patterns during shoot development in *Arabidopsis thaliana*. Plant J 32:867–877

Casson S, Spencer M, Walker K, Lindsey K (2005) Laser capture microdissection for the analysis of gene expression during embryogenesis of *Arabidopsis*. Plant J 42:111–123

Charrière F, Hahne G (1998) Induction of embryogenesis versus caulogenesis on in vitro cultured sunflower (*Helianthus annuus* L.) immature zygotic embryos: role of plant growth regulators. Plant Sci 137:63–71

Chen L et al (2011) The role of WRKY transcription factors in plant abiotic stresses. Biochim Biophys Acta. doi:10.1016/j.bbagrm.2011.09.002

Cheong Y, Chang H, Gupta R, Wang X, Zhu T, Luan S (2002) Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*. Plant Physiol 129:661–677

Chugh A, Khurana P (2002) Gene expression during somatic embryogenesis-recent advances. Curr Sci 83:715–730

Cueva Agila AY, Guachizaca I, Cella R (2013) Combination of 2,4-D and stress improves indirect somatic embryogenesis in *Cattleya maxima* Lindl. Plant Biosyst – Int J Dealing Asp Plant Biol 1–7

Cueva A, Concia L, Cella R (2012) Molecular characterization of a *Cyrtochilum loxense* somatic embryogenesis receptor-like kinase (SERK) gene expressed during somatic embryogenesis. Plant Cell Rep 31:1129–1139

Cui KR, Pei XW, Qin L, Wang JJ, Wang YF (1998) Effects of modulation of abscisic acid during somatic embryogenesis in *Lycium barbarum* L. Shi Yan Sheng Wu Xue Bao 31:195–201 (in Chinese)

de Vries SC, Booij H, Meyerink P, Huisman G, Wilde DH, Thomas TL et al (1988) Acquisition of embryogenic potential in carrot cell suspension cultures. Planta 176:196–204

Dodeman V, Ducreux G, Kreis M (1997) Zygotic embryogenesis versus somatic embryogenesis. J Exp Bot 48:1493–1509

Du H, Liu H, Xiong L (2013) Endogenous auxin and jasmonic acid levels are differentially modulated by abiotic stresses in rice. Front Plant Physiol 4:397

Dubois T, Guedira M, Dubois J, Vasseur J (1991) Direct somatic embryogenesis in leaves of *Cichorium*. Protoplasma 162:120–127

Dure L, Crouch M, Harada J, Ho T, Mundy J, Quatrano R et al (1989) Common amino acid sequence domains among the LEA proteins of higher plants. Plant Mol Biol 12:475–486

Feher A, Pasternak TP, Dudits D (2003) Transition of somatic plant cells to an embryogenic state. Plant Cell Tiss Org Cult 74:201–228

Frederico AM, Campos MD, Cardoso HG, Imani J, Arnholdt-Schmitt B (2009) Alternative oxidase involvement in *Dacus carota* somatic embryogenesis. Physiol Plant 137:498–508

Fujimura T, Komamine A (1975) Effects of various growth regulators on the embryogenesis in a carrot cell suspension culture. Plant Sci Lett 5:359–364

Fujimura T, Komamine A (1980) Mode of action of 2,4-D and zeatin on somatic embryogenesis in a carrot cell suspension culture. Z Pflanzenphysiol 99:1–8

Fujimura T, Komaminea A (1979) Involvement of endogenous auxin in somatic embryogenesis in a carrot cell suspension culture. Z Pflanzenphysiol 95:13–19

Gall OL, Torregrosa L, Danglot Y, Candresse T, Bouquet A (1994) Agrobacterium-mediated genetic transformation of grapevine somatic embryos and regeneration of transgenic plants expressing the coat protein of grapevine chrome mosaic nepovirus (GCMV). Plant Sci 102:161–170

Galland R, Randoux B, Vasseur J, Hilbert J (2001) Glutathione S transferase cDNA identified by mRNA differential display is upregulated during somatic embryogenesis in *Cichorium*. Biochim Biophys Acta 1522:212–216

Gavish H, Vardi A, Fluhr R (1991) Extra-cellular proteins and early embryo development in Citrus nucellar cell cultures. Physiol Plant 82:601–616

Ge X-X, Chai L-J, Liu Z, Wu X-M, Deng X-X, Guo W-W (2012) Transcriptional profiling of genes involved in embryogenic, non-embryogenic calluses and somatic embryogenesis of Valencia sweet orange by SSH based microarray. Planta 236:1107–1124

Gordon S, Heisler M, Reddy G, Ohno C, Das P, Meyerowitz E (2007) Pattern formation during de novo assembly of the Arabidopsis shoot meristem. Development 134:3539–3548

Gray DJ, Conger BV, Hanning GE (1984) Somatic embryogenesis in suspension and suspension-derived callus cultures of *Dactylis glomerata*. Protoplasma 122:196–202

Gray DJ, Trigiano RN, Conger BV (1993) Liquid suspension culture production of orchardgrass somatic embryos and their potential for the breeding of improved varieties. CRC Press, Boca Raton

Gribaudo I, Gambino G, Vallania R (2004) Somatic embryogenesis from grapevine anthers: identification of the optimal developmental stage for collecting explants. Am J Enol Vitic 55:427–430

Guiderdoni E, Mérot B, Eksomtramage T, Paulet F, Feldmann P, Glaszmann JC (1995) Somatic embryogenesis in sugarcane (*Saccharum* species). In: Bajaj YPS (ed) Somatic embryogenesis and synthetic seed II. Springer, Berlin, pp 92–113

Hanning GE, Conger BV (1982) Embryoid and plantlet formation from leaf segments of *Dactylis glomerata* L. Theor Appl Genet 63:155–159

Hatanaka T, Sawabe E, Azuma T, Uchida N, Yasuda T (1995) The role of ethylene in somatic embryogenesis from leaf discs of *Coffea canephora*. Plant Sci 107:199–204

Hatzopoulos P, Fong F, Sung Z (1990) Abscisic acid regulation of DC8, a carrot embryonic gene. Plant Physiol 94:690–695

Hecht V, Vielle-Calzada J, Hartog M, Schmidt E, Boutilier K, Grossniklaus U et al (2001) The *Arabidopsis* somatic embryogenesis receptor kinase 1 gene is expressed in developing ovules and embryos and enhances embryogenic competence in culture. Plant Physiol 127:803–816

Heidmann I, Lambalk J, Joosen R, Angenent G, Custers J, Boutilier K (2006) Expression of BABY BOOM induces somatic embryogenesis in tobacco. Int Conf "Haploids in Higher Plants III", Vienna, Austria, vol 52, pp 12–15

Hita O, Lafarga C, Guerra H (1997) Somatic embryogenesis from chickpea (*Cicer arietinum* L.) immature cotyledons: the effect of zeatin, gibberellic acid and indole-3-butyric acid. Acta Physiol Plant 19:333–338

Ho W-J, Vasil IK (1983) Somatic embryogenesis in sugarcane (*Saccharum officinarum* L.). I. The morphology and physiology of callus formation and the ontogeny of somatic embryos. Protoplasma 118:169–180

Horstman A, Fukuoka H, Weemen M, Angenent G, Fiers M, Boutilier K (2009). Signalling components of BABY BOOM-induced somatic embryogenesis. Plant Research International, Wageningen, The Netherlands, National Institute of Vegetable and Tea Science, Mie, Japan

Hu H, Xiong L, Yang Y (2005) Rice SERK1 gene positively regulates somatic embryogenesis of cultured cell and host defense response against fungal infection. Planta 222:107–117

Huang X, Li X, Li Y, Huang L (2001) The effect of AOA on ethylene and polyamine metabolism during early phases of somatic embryogenesis in *Medicago sativa*. Physiol Plant 113: 424–429

Hughes D, Galau G (1991) Developmental and environmental induction of Lea and LeaA mRNAs and the postabscission program during embryo culture. Plant Cell 3:605–618

Hutchinson MJ, KrishnaRaj S, Saxena PK (1997) Inhibitory effect of GA3 on the development of thidiazuron-induced somatic embryogenesis in geranium (*Pelargonium·hortorum* Bailey) hypocotyl cultures. Plant Cell Rep 16:435–438

Iantcheva A, Vlahova M, Bakalova E, Kondorosi E, Elliott MC, Atanassov A (1999) Regeneration of diploid annual medics via direct somatic embryogenesis promoted by thidiazuron and benzylaminopurine. Plant Cell Rep 18:904–910

Ikeda M, Umehara M, Kamada H (2006) Embryogenesis-related genes; its expression and roles during somatic and zygotic embryogenesis in carrot and Arabidopsis. Plant Biotechnol 23:153–161

Indra AP, Krishnaveni S (2009) Effect of hormones, explants and genotypes in in vitro culturing of sorghum. J Biochem Technol 1:96–103

Ito Y, Takaya K, Kurata N (2005) Expression of SERK family receptor-like protein kinase genes in rice. Biochim Biophys Acta 1730:253–258

Ivanova A, Velcheva M, Denchev P, Atanassov A, Van OH (1994) Endogenous hormone levels during direct somatic embryogenesis in *Medicago falcata*. Physiol Plant 92:85–89

Iwai M, Umehara M, Satoh S, Kamada H (2003) Stress-induced somatic embryogenesis in vegetative tissues of *Arabidopsis thaliana*. Plant J 34:107–114

Jackson C, Casanova J (2000) Turning on ARF: the Sec7 family of guanine-nucleotide-exchange factors. Trends Cell Biol 10:60–67

Jain M, Khurana JP (2009) Transcript profiling reveals diverse roles of auxin-responsive genes during reproductive development and abiotic stress in rice. FEBS J 276:3148–3162

Jansen M, Booij H, Schel J, de Vries S (1990) Calcium increases the yield of somatic embryos in carrot embryogenic suspension cultures. Plant Cell Rep 9:221–223

Jiménez VM (2001) Regulation of in vitro somatic embryogenesis with emphasis on the role of endogenous hormones. Revista Brasileira de Fisiologia Vegetal 13(2):196–223

Jiménez VM (2005) Involvement of plant hormones and plant growth regulators on in vitro somatic embryogenesis. Plant Growth Regul 47:91–110

Jimenez VM, Bangerth F (2001) Hormonal status of maize initial explants and of the embryogenic and non-embryogenic callus cultures derived from them as related to morphogenesis in vitro. Plant Sci 160(2):247–257

Jiménez V, Bangerth F (2001a) Endogenous hormone levels in explants and in embryogenic and non-embryogenic cultures of carrot. Physiol Plant 111:389–395

Jiménez VM, Bangerth F (2001b) Endogenous hormone levels in initial explants and in embryogenic and non-embryogenic callus cultures of competent and non-competent wheat genotypes. Plant Cell Tiss Org Cult 67:37–46

Jones TJ, Rost TL (1989) The developmental anatomy and ultrastructural of somatic embryos from rice (*Oryza sativa* L.) scutellum epithelial cells. Bot Gaz 150:41–49

Kairong C, Gengsheng X, Xinmin L, Gengmei X, Yafu W (1999) Effect of hydrogen peroxide on somatic embryogenesis of *Lycium barbarum* L. Plant Sci 146:9–16

Kamada H, Kobayashi K, Kiyosue T, Hiroshi H (1989) Stress induced somatic embryogenesis in carrot and its application to synthetic seed production. In Vitro Cell Dev Biol 25:1163–1166

Kamada H, Ishikawa K, Saga H, Harada H (1993) Induction of somatic embryogenesis in carrot by osmotic stress. Plant Tissue Cult Lett 10:38–44

Kamada H, Tachikawa Y, Saitou T, Harada H (1994) Heat stresses induction of carrot somatic embryogenesis. Plant Tissue Cult Lett 11:229–232

Karami O, Deljou A, Esna-Ashari M, Ostad-Ahmadi P (2006) Effect of sucrose concentrations on somatic embryogenesis in carnation (*Dianthus caryophyllus* L.). Sci Hortic 110:340–344

Kasajima I, Ide Y, Hirai MY, Fujiwara T (2010) WRKY6 is involved in the response to boron deficiency in *Arabidopsis thaliana*. Physiol Plant 139:80–92

Kawahara R, Komamine A (1995) Molecular basis of somatic embryogenesis. In: Bajaj Y (ed) Biotechnology in agriculture and forestry, somatic embryogenesis and synthetic seed. Springer, Berlin, pp 30–40

Kepczynski J, Mckersie BD, Brown DCW (1992) Requirement of ethylene for growth of callus and somatic embryogenesis in *Medicago sativa* L. J Exp Bot 43:1199–1202

Kiyosue T, Yamaguchi-Shinozaki K, Shinozaki K, Higashi K, Satoh S, Kamada H et al (1992) Isolation and characterization of a cDNA that encodes ECP31, an embryogenic-cell protein from carrot. Plant Mol Biol 19:239–249

Kiyosue T, Satoh S, Kamada H, Harada H (1993a) Somatic embryogenesis in higher plants. J Plant Res 3:75–82

Kiyosue T, Yamaguchi-Shinozaki K, Shinozaki K, Kamada H, Harada H (1993b) cDNA cloning of ECP40, an embryogenic-cell protein in carrot, and its expression during somatic and zygotic embryogenesis. Plant Mol Biol 21(6):1053–1068

Kobayashi T, Higashi K, Sasaki K, Asami T, Yoshida S, Kamada H (2000) Purification from conditioned medium and chemical identification of a factor that inhibits somatic embryogenesis in carrot. Plant Cell Physiol 41:268–273

Kochba J, Spiegel-Roy P, Neumann H, Saad S (1978) Stimulation of embryogenesis in Citrus ovular callus by ABA, Ethephon, CCC and Alar and its suppression by GA3. Z Pflanzenphysiol 89:427–432

Kong L, Von Aderkas P (2007) Genotype effects on ABA consumption and somatic embryo maturation in interior spruce (*Picea glauca*×*engelmanni*). J Exp Bot 58:1525–1531

Kong L, Yeung E (1994) Effects of ethylene and ethylene inhibitors on white spruce somatic embryo maturation. Plant Sci 104:71–80

Kong L, Attree S, Evans D, Binarova P, Yeung E, Fowke L (1999) Somatic embryogenesis in white spruce: studies of embryo development and cell biology. In: Jain SM, Gupta PK, Newton RJ (eds) Somatic embryogenesis in woody plants. Kluwer Academic Publishers, Dordrecht, pp 1–28

Kulinska-Lukaszek K, Tobojka M, Adamiok A, Kurczynska EU (2012) Expression of the BBM gene during somatic embryogenesis of *Arabidopsis thaliana*. Biol Plant 56:389–394

Kumar PP, Joy RW IV, Thorpe TA (1989) Ethylene and carbon dioxide accumulation, and growth of cell suspension cultures of *Picea glauca* (white spruce). J Plant Physiol 135:592–596

Kumria R, Sunnichan V, Das D, Gupta S, Reddy V, Bhatnagar R et al (2003) High-frequency somatic embryo production and maturation into normal plants in cotton (*Gossypium hirsutum*) through metabolic stress. Plant Cell Rep 21:635–639

Lagace M, Matton DP (2004) Characterization of a WRKY transcription factor expressed in late torpedo-stage embryos of *Solanum chacoense*. Planta 219:185–189

Lee E, Cho D, Soh W (2001) Enhanced production and germination of somatic embryos by temporary starvation in tissue cultures of *Daucus carota*. Plant Cell Rep 20:408–415

Li T, Neumann KH (1985) Embryogenesis and endogenous hormone content of cell cultures of some carrot varieties (*Daucus carota* L.). Ber Deutsch Bot Ges 98:227–235

Li T, Chen J, Qiu S, Zhang Y, Wang P, Yang L et al (2012) Deep sequencing and microarray hybridization identify conserved and species-specific microRNAs during somatic embryogenesis in hybrid yellow poplar. PLoS One 7. doi:10.1371/journal.pone.0043451

Liu QC, Kokubu T, Sato M (1992) Varietal differences of somatic embryogenesis in shoot tip culture of sweetpotato, *Ipomoea batatas* (L.) Lam. Jpn J Breed 42:8–9

Liu CM, Xu ZH, Chua NH (1993) Auxin polar transport is essential for the establishment of bilateral symmetry during early plant embryogenesis. Plant Cell Rep 5:621–630

Maeda E, Radi SH (1991) Ultrastructural aspects of rice scutellum as related to seminal root cultures. Biotechnol Agric For (Rice) 14:78–91

Maheswaran G, Williams EG (1985) Origin and development of somatic embryoids formed directly on immature embryos of *Trifolium repens* in vitro. Ann Bot 56:619–630

Maillot P, Kieffer F, Walter B (2006) Somatic embryogenesis from stem nodal sections of grapevine. Vitis 45:185–189

Maillot P, Lebel S, Schellenbaum P, Jacques A, Walter B (2009) Differential regulation of SERK, LEC1-like and pathogenesis-related genes during indirect secondary somatic embryogenesis in grapevine. Plant Physiol Biochem 47:743–752

Mantiri FR, Kurdyukov S, Lohar DP, Sharopova N, Saeed NA, Wang X-D et al (2008) The transcription factor MtSERF1 of the ERF subfamily identified by transcriptional profiling is required for somatic embryogenesis induced by auxin plus cytokinin in *Medicago truncatula*. Plant Physiol 146:1622–1636

Marcotte WR Jr, Bayley CC, Quatrano RS (1988) Regulation of a wheat promoter by abscisic acid in rice protoplasts. Nature 335:454–457

Martin AB, Cuadrado Y, Guerra H, Gallego P, Hita O, Martin L et al (2000) Differences in the contents of total sugars, reducing sugars, starch and sucrose in embryogenic and non-embryogenic calli from *Medicago arborea* L. Plant Sci 154:143–151

Maxwell DP, Wang Y, McIntosh L (1999) The alternative oxidase lowers mitochondrial reactive oxygen production in plant cells. Proc Natl Acad Sci U S A 96:8271–8276

McCabe PF, Valentine TA, Forsberg LS, Pennell RI (1997) Soluble signals from cells identified at the cell wall establish a developmental pathway in carrot. Plant Cell 9:2225–2241

Michalczuk L, Druart P (1999) Indole-3-acetic acid metabolism in hormone-autotrophic, embryogenic callus of Inmil cherry rootstock (*Prunus incisa serrula* "GM 9") and in hormone-dependent, non-embryogenic calli of *Prunus incisa serrula* and *Prunus domestica*. Physiol Plant 107:426–443

Michalczuk L, Cooke T, Cohen J (1992) Auxin levels at different stages of carrot somatic embryogenesis. Phytochemistry 31:1097–1103

Mikuła A, Wilbik W, Rybczyński JJ (1998) Wplyw regulatoròw wzrostu na somatyczna embriogeneze Gentiana sp w kulturach in vitro. II Ogòlnopolska Konferencja Zastosowanie kultur in vitro wfizjologii roslin 141–154

Minocha R, Minocha SC, Long S (2004) Polyamines and their biosynthetic enzymes during somatic embryo development in red spruce (*Picea rubens* Sarg.). In Vitro Cell Dev Biol Plant 40:572–580

Mishra A, Khurana P (2003) Genotype dependent somatic embryogenesis and regeneration from leaf base cultures of *Sorghum bicolor*. J Plant Biochem Biotechnol 12:53–56

Moghaieb REA, Youssef SS, Mohammed EHK, Draz AE-SE (2009) Genotype dependent somatic embryogenesis from Egyptian rice mature zygotic embryos. Aust J Basic Appl Sci 3:2570–2580

Nagata T, Ishida S, Hasezawa S, Takahashi Y (1994) Genes involved in the dedifferentiation of plant cells. Int J Dev Biol 38:321–327

Namasivayam P (2007) Acquisition of embryogenic competence during somatic embryogenesis. Plant Cell Tiss Org Cult 90:1–8

Nissen P (1994) Stimulation of somatic embryogenesis in carrot by ethylene: effects of modulators of ethylene biosynthesis and action. Physiol Plant 92:397–403

Noctor G, Foyer C (1998) Ascorbate and glutathione: keeping active oxygen under control. Annu Rev Plant Physiol Plant Mol Biol 49:249–279

Nolan KE, Irwanto RR, Rose RJ (2003) Auxin up-regulates MtSERK1 expression in both *Medicago truncatula* root-forming and embryogenic cultures. Plant Physiol 133:218–230

Nolan K, Saeed N, Rose R (2006) The stress kinase gene MtSK1 in *Medicago truncatula* with particular reference to somatic embryogenesis. Plant Cell Rep 25:711–722

Nolan KE, Kurdyukov S, Rose RJ (2009) Expression of the SOMATIC EMBRYOGENESIS RECEPTOR-LIKE KINASE1 (SERK1) gene is associated with developmental change in the life cycle of the model legume *Medicago truncatula*. J Exp Bot 60:1759–1771

Nole-Wilson S, Tranby TL, Krizek BA (2005) AINTEGUMENTA-like (AIL) genes are expressed in young tissues and may specify meristematic or division-competent states. Plant Mol Biol 57:613–628

Nomura K, Komamine A (1985) Identification and isolation of single cells that produce somatic embryos at a high frequency in a carrot cell suspension culture. Plant Physiol 79:988–991

Okushima Y, Overvoordea PJ, Arimaa K, Alonsob JM, Chana A, Changa C et al (2005) Functional genomic analysis of the AUXIN RESPONSE FACTOR gene family members in *Arabidopsis thaliana*: unique and overlapping functions of ARF7 and ARF19. Plant Cell Physiol 17:444–463

Omid K, Abbas S (2010) The molecular basis for stress-induced acquisition of somatic embryogenesis. Mol Biol Rep 37:2493–2507

Ozias-Akins P, Anderson WF, Holbrook CC (1992) Somatic embryogenesis in *Arachis hypogaea* L.: genotype comparison. Plant Sci 83:103–111

Pandey DK, Singh AK, Chaudhary B (2012) Boron-mediated plant somatic embryogenesis: a provocative model. J Bot 2012. doi:10.1155/2012/375829

Pandey, D. K. and Chaudhary, B. (2014) Oxidative stress responsive *SERK1* gene directs the progression of somatic embryogenesis in cotton (*Gossypium hirsutum* L. cv. Coker 310). American J Plant Sci 5:80-102

Passarinho P, Ketelaar T, Xing M, Van Arkel J, Maliepaard C, Hendriks M et al (2008) BABY BOOM target genes provide diverse entry points into cell proliferation and cell growth pathways. Plant Mol Biol 68:225–237

Patnaik D, Mahalakshmi A, Khurana P (2005) Effect of water stress and heavy metals on induction of somatic embryogenesis in wheat leaf base cultures. Indian J Exp Biol 43:740–745

Perez-Nunez M, Souza R, Seaenz L, Chan J, Zuniga-Aguilar J, Oropeza C (2009) Detection of a SERK-like gene in coconut and analysis of its expression during the formation of embryogenic callus and somatic embryos. Plant Cell Rep 28:11–19

Perrin M, Gertz C, Masson J (2004) High efficiency initiation of regenerable embryonic callus from anther filaments of 19-grapevine genotypes grown worldwide. Plant Sci 167:1343–1349

Ptak A, Tahchy AE, Wyzgolik G, Henry M, Laurain-Mattar D (2010) Effects of ethylene on somatic embryogenesis and galanthamine content in *Leucojum aestivum* L. cultures. Plant Cell Tiss Org Cult 102:61–67

Quiroz-Figueroa F, Rojas-Herrera R, Galaz-Avolos R, Loyola-Vargas V (2006) Embryo production through somatic embryogenesis can be used to study cell differentiation in plants. Plant Cell Tiss Org Cult 86:285–301

Radojevic L (1979) Somatic embryos and plantlets from callus cultures of *Paulownia tomentosa* Steud. Z Pflanzenphysiol 9:57–62

Rajasekaran K, Hein MB, Vasil IK (1987) Endogenous abscisic acid and indole-3-acetic acid and somatic embryogenesis in cultured leaf explants of *Pennisetum purpureum* Schum. Plant Physiol 84:47–51

Rao AQ, Hussain SS, Shahzad MS, Bokhari SYA, Raza MH, Rakha A et al (2006) Somatic embryogenesis in wild relatives of cotton (*Gossypium* spp.). J Zhejiang Univ Sci B 7:291–298

Remotti PC (1995) Primary and secondary embryogenesis from cell suspension cultures of *Gladiolus*. Plant Sci 107:205–214

Robertson D, Weissinger AK, Ackley R, Glover S, Sederoff RR (1992) Genetic transformation of Norway spruce (*Picea abies* (L.) Karst) using somatic embryo explants by microprojectile bombardment. Plant Mol Biol 19:925–935

Rose RJ, Nolan KE (2006) Genetic regulation of somatic embryogenesis with particular reference to *Arabidopsis thaliana* and *Medicago truncatula*. In Vitro Cell Dev Biol 42:473–481

Roustan J-P, Latché A, Fallot J (1990) Control of carrot somatic embryogenesis by $AgNO_3$, an inhibitor of ethylene action: effect on arginine decarboxylase activity. Plant Sci 67:89–95

Ruduś I, Kępczyńska E, Kępczyński J (2002) Regulation of *Medicago sativa* L. somatic embryogenesis by gibberellins. Plant Growth Regulat 36:91–95

Sakhanokho HF, Ozias-Akins P, May OL, Chee PW (2004) Induction of somatic embryogenesis and plant regeneration in select Georgia and Pee Dee cotton lines. Crop Sci 44:2199–2205

Santa-Catarina C, Oliveira RRd, Cutri L, Floh EIS, Dornelas MC (2012) WUSCHEL-related genes are expressed during somatic embryogenesis of the basal angiosperm *Ocotea catharinensis* Mez. (Lauraceae). Trees 26:493–501

Santarem E, Pelissier B, Finer J (1997) Effect of explant orientation, pH, solidifying agent and wounding on initiation of soybean somatic embryos. In Vitro Cell Dev Biol Plant 33:13–19

Santos M, Romano E, Yotoko K, Tinoco M, Dias B, Argao F (2005) Characterisation of the cacao somatic embryogenesis receptor-like kinase gene expressed during somatic embryogenesis. Plant Sci 168:723–729

Sasaki K, Shimomura K, Kamada H, Harada H (1994) IAA metabolism in embryogenic and non-embryogenic carrot cells. Plant Cell Physiol 35:1159–1164

Schiavone FM, Cooke TJ (1985) A geometric analysis of somatic embryo formation in carrot cell cultures. Can J Bot 63:1573–1578

Schiavone FM, Cooke TJ (1987) Unusual patterns of somatic embryogenesis in the domesticated carrot: developmental effects of exogenous auxins and auxin transport inhibitors. Cell Differ 21:53–62

Schlögl PS, Santos ALWd, Vieira LdN, Floh EIS, Guerra MP (2012) Gene expression during early somatic embryogenesis in Brazilian pine (*Araucaria angustifolia* (Bert) O. Ktze). Plant Cell Tiss Org Cult 108:173–180

Schmidt E, Guzzo F, Toonen M, De Vries S (1997) A leucine-rich repeat containing receptor-like kinase marks somatic plant cells competent to form embryos. Development 124:2049–2062

Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y et al (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. Plant J 31:279–292

Sharma S, Millam S, Hein I, Bryan G (2008) Cloning and molecular characterisation of a potato SERK gene transcriptionally induced during initiation of somatic embryogenesis. Planta 228:319–330

Shimada T, Hirabayashi T, Fujii H, Kita M, Omura M (2005) Isolation and characterization of the somatic embryogenesis receptor-like kinase gene homologue (CitSERK) from *Citrus unshiu* Marc. Sci Hort 103:233–238

Singla B, Khurana JP, Khurana P (2008) Characterization of three somatic embryogenesis receptor kinase genes from wheat, *Triticum aestivum*. Plant Cell Rep 27:833–843

Srinivasan C, Liu Z, Heidmann I, Supena E, Fukuoka H, Joosen R et al (2007) Heterologous expression of the BABY BOOM AP2/ERF transcription factor enhances the regeneration capacity of tobacco (*Nicotiana tabacum* L.). Planta 225:341–351

Steenhoudt O, Vanderleyden J (2000) Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects. FEMS Microbiol Rev 24:487–506

Steiner N, Santa-Catarina C, Guerra MP, Cutri L, Dornelas MC, Floh EIS (2012) A gymnosperm homolog of somatic embryo receptor like kinase (SERK1) is expressed during somatic embryogenesis. Plant Cell Tiss Org Cult 109:41–50

Steward F, Mapes M, Hears K (1958) Growth and organized development of cultured cells. II. Growth and division of freely suspended cells. Am J Bot 45:705–708

Su Y, Zhao X, Liu Y, Zhang C, O'Neill S, Zhang X (2009) Auxin-induced WUS expression is essential for embryonic stem cell renewal during somatic embryogenesis in Arabidopsis. Plant J 59:448–460

Takano J, Wada M, Ludewig U, Schaaf G, von Wirén N, Fujiwara T (2006) The *Arabidopsis* major intrinsic protein NIP5;1 is essential for efficient boron uptake and plant development under boron limitation. Plant Cell 18:1498–1509

Takeda T, Inose H, Matsuoka H (2003) Stimulation of somatic embryogenesis in carrot cells by the addition of calcium. Biochem Eng J 14:143–148

Tereso S, Zoglauer K, Milhinhos A, Miguel C, Oliveira M (2007) Zygotic and somatic embryo morphogenesis in *Pinus pinaster*: comparative histological and histochemical study. Tree Physiol 27:661–669

Thibaud-Nissen F, Shealy R, Khanna A, Vodkin L (2003) Clustering of microarray data reveals transcript patterns associated with somatic embryogenesis in soybean. Plant Physiol 132:118–136

Thomas C, Meyer D, Himber C, Steinmetz A (2004) Spatial expression of a sunflower SERK gene during induction of somatic embryogenesis and shoot organogenesis. Plant Physiol Biochem 42:35–42

Tokuji Y, Kuriyama KJ (2003) Involvement of gibberellin and cytokinin in the formation of embryogenic cell clumps in carrot (*Daucus carota*). Plant Physiol 160:133–141

Toonen MAJ, De Vries SC (1996) Initiation of somatic embryos from single cells. In: Wang TL, Cuming A (eds) Embryogenesis: the generation of a plant. Bios Scientific Publishers, Oxford, pp 173–189

Touraev A, Vicente O, Heberlebors E (1997) Initiation of microspore embryogenesis by stress. Trends Plant Sci 2:297–302

Trolinder N, Goodin J (1987) Somatic embryogenesis and plant regeneration in cotton (*Gossypium hirsutum* L.). Plant Cell Rep 6:231–234

Trolinder NL, Xhixian C (1989) Genotype specificity of the somatic embryogenesis response in cotton. Plant Cell Rep 8:133–136

Tsuchisaka A, Yu G, Jin H, Alonso J, Ecker J, Zhang X et al (2009) A combinatorial interplay among the 1-aminocyclopropane-1-carboxylate isoforms regulates ethylene biosynthesis in *Arabidopsis thaliana*. Genetics 183:979–1003

Vahdati K, Bayat S, Ebrahimzadeh H, Jariteh M, Mirmasoumi M (2008) Effect of exogenous ABA on somatic embryo maturation and germination in Persian walnut (*Juglans regia* L.). Plant Cell Tiss Org Cult 93:163–171

Vain P, Flament P, Soudain P (1989) Role of ethylene in embryogenic callus initiation and regeneration in *Zea mays* L. J Plant Physiol 135:537–540

Wang D, Pei K, Fu Y, Sun Z, Li S, Liu H et al (2007) Genome-wide analysis of the auxin response factors (ARF) gene family in rice (*Oryza sativa*). Gene 394:13–24

Warren WJ, Warren WP (1993) Mechanisms of auxin regulation of structural and physiological polarity in plants, tissues, cells and embryos. Aust J Plant Physiol 20:555–571

Wei Tan (2001) Somatic embryogenesis and peroxidase activity of desiccation tolerant mature somatic embryos of loblolly pine. J Forestry Res 12(3):147–152

Willekens H, Chamnongpol S, Davey M, Schraudner M, Langebartels C, Montagu MV et al (1997) Catalase is a sink for H_2O_2 and is indispensable for stress defence in C_3 plants. EMBO J 16:4806–4816

Wiśniewskaa A, Grabowskab A, Pietraszewska-Bogielc A, Tagashirad N, Zuzgac S, Wóycickic R et al (2012) Identification of genes up-regulated during somatic embryogenesis of cucumber. Plant Physiol Biochem 50:54–64

Wójcikowska B, Jaskóła K, Gąsiorek P, Meus M, Nowak K, Gaj MD (2013) LEAFY COTYLEDON2 (LEC2) promotes embryogenic induction in somatic tissues of Arabidopsis, via YUCCA-mediated auxin biosynthesis. Planta 1–16. doi:10.1007/s00425-013-1892-2

Xu Z, Zhang C, Zhang X, Liu C, Wu Z, Yang Z et al (2013) Transcriptome profiling reveals auxin and cytokinin regulating somatic embryogenesis in different sister lines of cotton cultivar CCRI24. J Integr Plant Biol 55:631–642

Yang H, Saitou T, Komeda Y, Harada H, Kamada H (1996) Late embryogenesis abundant protein in *Arabidopsis thaliana* homologous to carrot ECP31. Physiol Plant 98:661–666

Yang H, Saitou T, Komeda Y, Harada H, Kamada H (1997) *Arabidopsis thaliana* ECP63 encoding a LEA protein is located in chromosome 4. Gene 184:83–88

Yang X, Zhang X, Yuan D, Jin F, Zhang Y, Xu J (2012) Transcript profiling reveals complex auxin signalling pathway and transcription regulation involved in dedifferentiation and redifferentiation during somatic embryogenesis in cotton. BMC Plant Biol 12. doi:10.1186/471-2229-12-110

Yang X, Wang L, Yuan D, Lindsey K, Zhang X (2013) Small RNA and degradome sequencing reveal complemiRNA regulation during cotton somatic embryogenesis. J Exp Bot 64:1521–1536

Yasuda T, Fujii Y, Yamaguchi T (1985) Embryogenic callus induction from *Coffea arabica* leaf explants by benzyladenine. Plant Cell Physiol 26:595–597

Zdravković-Korać S, Nešković M (1999) Induction and development of somatic embryos from spinach (*Spinacia oleracea*) leaf segments. Plant Cell Tiss Org Cult 55:109–114

Zhang C-X, Yao Z-Y, Zhao Z, Qi J-H (1997) Histological observation of somatic embryogenesis from cultured embryos of *Quercus variabilis* B1. J Plant Physiol Mol Biol 33:33–38

Zhang B-H, Feng R, Liu F, Wang Q (2001) High frequency somatic embryogenesis and plant regeneration of an elite Chinese cotton variety. Bot Bull Acad Sin 42:9–16

Zheng Q, Zheng Y, Perry SE (2013) AGAMOUS-Like15 promotes somatic embryogenesis in Arabidopsis and soybean in part by the control of ethylene biosynthesis and response. Plant Physiol 161:2113–2127

Zhenga Y, Renb N, Wanga H, Strombergb AJ, Perrya SE (2009) Global identification of targets of the *Arabidopsis* MADS domain protein AGAMOUS-Like15. Plant Cell 21:2563–2577

Zou X, Seemann JR, Neuman D, Shen QJ (2004) A WRKY gene from creosote bush encodes an activator of the abscisic acid signaling pathway. J Biol Chem 279:55770–55779

Zuo J, Niu Q-W, Frugis G, Chua N-H (2002) The WUSCHEL gene promotes vegetative-to-embryonic transition in Arabidopsis. Plant J 30:349–359

Chapter 11
Bioinformatics Tools to Analyze Proteome and Genome Data

Ritesh Kumar, Shalini Singh, and Vikash Kumar Dubey

Introduction

Genomics is study of all genes present in a living system. Gene makes pre mRNA that may undergo several alternate splicing making various possible mRNA that translate into protein. Thus, a single gene may code for several proteins. Proteins may also undergo different post-translational modifications. It is easily understandable that the Proteome (complete set of protein in a living organism) is much more complex compared to genomics. Biological system involves complex function of proteome and genome. Recent development in science has generated vast amount of proteomics and genomics data that has to be managed and analyzed for useful information to understand biological systems (Fig. 11.1). Such data generated by high-throughput methods are stored in public repositories like NCBI GEO (Barrett et al. 2005), ArrayExpress (Parkinson et al. 2005), UniPROBE (Newburger and Bulyk 2009), SWISS-2DPAGE and Two-dimensional polyacrylamide gel electrophoresis database. Several such proteome and genome public repositories exist. Bioinformatics has also made tremendous progress in last decade. Several tools have been developed to analyze and understand vast amount of proteome and genome data.

*Author contributed equally with all other contributors.

R. Kumar, Ph.D. (pursuing) • S. Singh, Ph.D. (pursuing) • V.K. Dubey, Ph.D.(✉)
Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati 781039, Assam, India
e-mail: vdubey@iitg.ernet.in

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_11

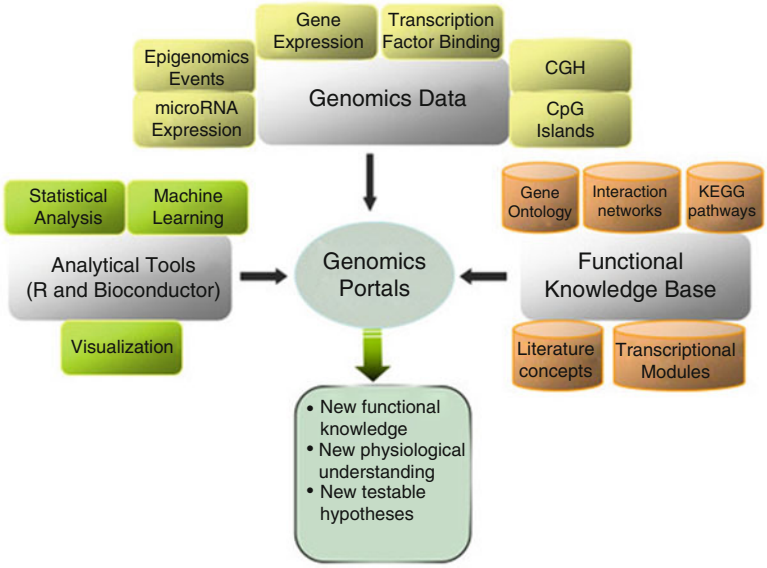

Fig. 11.1 Genomics portals. Genome datasets need to be analyzed for function, change in expression in different conditions, etc. (Figure adopted from *BMC Genomics* 2010, 11:27)

Bioinformatics Tool to Analyze Proteomics Data

Presently, there are numerous methods available for the analysis of proteins and peptides but in the beginning, 2D PAGE was one of the best techniques to resolve simultaneously thousands of proteins based on their charge (pI) and molecular weight. The basic workflow for the analysis of proteome by 2D PAGE and mass spectrometric analysis is shown in Fig. 11.2. The principal procedure of 2D PAGE is based on the isoelectric focusing (IEF) in first dimension, which separates proteins based on their isoelectric point followed by SDS-Polyacrylamide gel electrophoresis (SDS-PAGE) in second dimension that separates protein based on their molecular weights. These separation techniques allow complex mixtures of proteins to be separated differing by a single charge, thereby allowing any modifications to be detected. IEF is the most important step in 2D-PAGE, in which proteins are solubilized in high concentration of urea and reducing agents without SDS. It also requires the first dimension pH range based on protein sample and length of the strip based on size of SDS-PAGE.

At this point, movement of proteins depend on their net charge and pI. A protein with net negative charge will move towards positive electrode and protein with net positive charge will move towards negative electrode. The electrophoretic mobility of protein becomes zero when pH is equal to its pI. In second dimension of 2D-PAGE, separation of proteins is based on the difference in electrophoretic mobility due to difference in their molecular radius which is roughly equal to its size. Proteins are already denatured in first dimension by urea and reducing agents, therefore the SDS

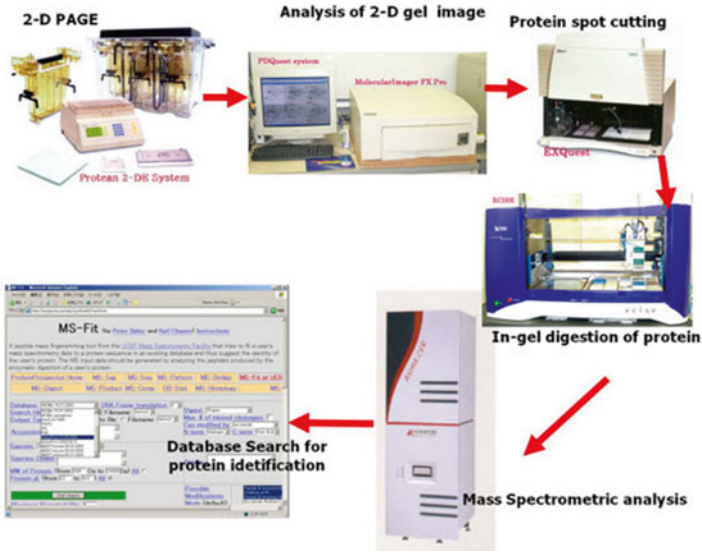

Fig. 11.2 Flow chart showing different steps in proteome analysis. After two-dimensional gel electrophoresis, differentially expressed protein spot is identified using image analysis. Subsequently, differentially expressed proteins are identified by in-gel digestion of protein spot by a protease and fragments mass determination by mass spectrometry. Finally, data is analyzed using computational tools. (Figure and part of legend adopted from *BMC Bioinformatics*, 2006, 7:430)

present in gel running buffer is sufficient to bind with already denatured proteins. Generally 1.4 g of SDS binds to 1 g of protein and forms uniformly negative charge complex that moves towards positively charged electrode (Reynolds and Tanford 1970). Certain protein contains phosphate, lipid, and carbohydrate groups and can bind to varying amount of SDS, resulting in abnormal electrophoretic mobility. After second dimension run, gel is separated and stained for protein visualization (Fig. 11.3). The high resolving capacity of 2D-Gel and due to various staining procedures available, it is very much useful to resolve thousand of proteins and to identify changes in protein abundance between two proteome samples.

Protein samples from control and treated cells are visualized by software systems for 2D-gel image analysis. The common software that are used frequently are: (1) Image master 2D-platinum and (2) DeCyder, both are commercially available on GE healthcare (www.gelifesciences.com) (3) PDQuest (4) Proteomweaver, are commercially available on Bio-Rad (www.bio-rad.com) and (5) Delta2D which is commercially available on Decodon (www.decodon.com). The common task for 2D gel users are to choose appropriate tools/softwares for their need. Most of the software systems for 2D gel image analysis have common function. These are visualization, quantification of protein spots on the gel, and matching of corresponding spots across the gel. Sometimes, more than one spot per protein as well as co-migrating spots are present on gel which creates problems for quantification of protein, database matching, and comparison of proteins across the gel. There are

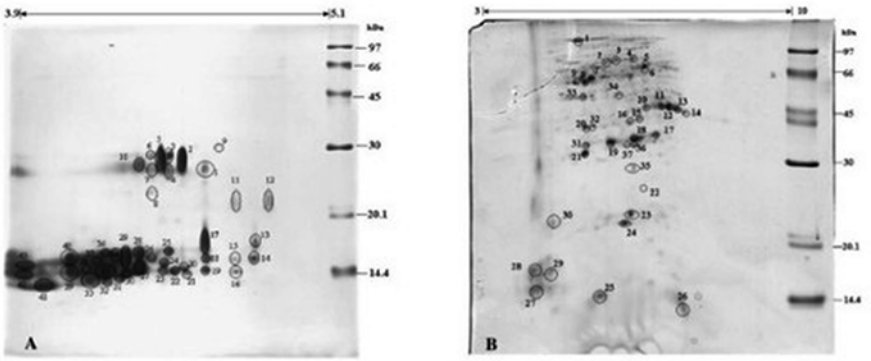

Fig. 11.3 A typical image of two-dimensional (IEF-SDS-PAGE) gel electrophoresis of crude phenol-soluble protein fraction isolated from *C. comosa* rhizomes using isoelectric focusing in pH range of (**a**) 3–5.4 and (**b**) 5.4–10. (Figure and part of legend adopted from *Proteome Science*, 2011, 9:43)

Table 11.1 Types of cyDye used for the labeling of protein samples in 2D-DIGE

Name of dye	Dyes commercially available	Amino acids labeled	Sensitivity
CyDye DIGE Fluor minimal dye	Cy2, Cy3, Cy5	A small % of lysine residues	Equivalent to silver staining
CyDye DIGE Fluor saturation dye	Cy3, Cy5	100 % of cystein residues	One hundred times higher than silver staining

Source: Adopted from *Methods in Molecular Biology*, 2009, 577)

various interfering substances in 2D gel electrophoresis that can interfere with separation and visualization of proteome. The most common non-protein impurities are (1) salt (2) lipids (3) nucleic acids and (4) carbohydrates. Salt contamination causes most frequent and insufficient focusing in first dimension IEF run, whereas lipids, charged polysaccharides, and nucleic acid binding proteins can bind to proteins changing both their isoelectric point and molecular weight.

In regular two-dimensional gel electrophoresis, control and treated protein samples run on two different gels which are visualized and compared by software systems as mentioned earlier. However, the results are always doubtful due to gel to gel variations. Differential imaging of gel electrophoresis (2D-DIGE) is one of the very effective tools that resolves and quantifies different protein samples on single gel. It uses fluorescence dye (cy2, cy3 and cy5). Protein samples are labeled with CyDye prior to IEF and SDS-PAGE (Table 11.1). There are two forms of CyDye labeling: (1) minimal labeling and (2) saturation labeling. In minimal labeling ε-amino group of lysine of protein samples covalently react with N-hydroxy succinamide ester group of CyDye. Concentration of dye is kept limiting to 2–3 % so that only a single lysine per protein molecule is labeled and rest remains unlabelled. Binding of these dyes does not affect the isoelectric point of protein because the fluor dye contains +1 charge but it adds 450 Da of mass. The detection limit is 100–200 pg of single

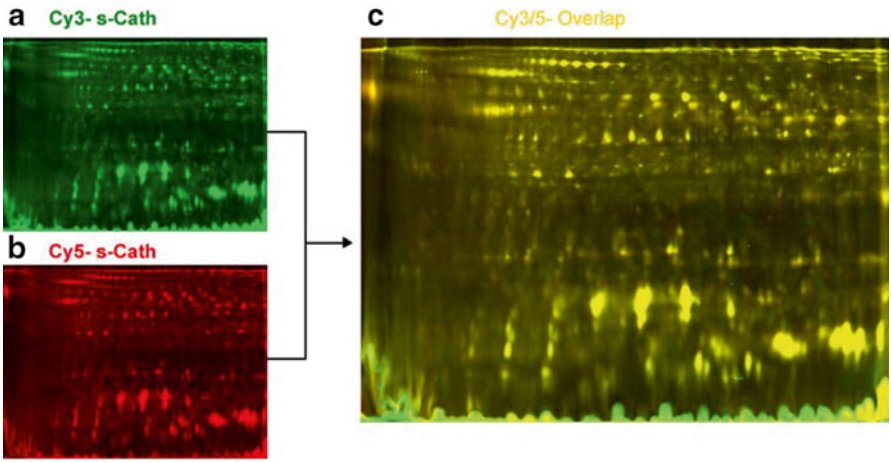

Fig. 11.4 Assessment of experimental variability by 2D-DIGE in proteomic analysis. Identical samples were labeled with **(a)** Cy3 and **(b)** Cy5, and run on a single gel. A total of 2,257 spots were detected in the overlapped image. **(c)** All 2,257 spots were found to be similar between the Cy3 and Cy5 images by the Decyder program. In the overlapped image, spots that are similar between Cy3 and Cy5 are *yellow*. Spots that are increased in Cy3 and in Cy5 are *green* and *red*, respectively. (Figure and figure legend are adopted from *Clinical Proteomics*, 2007, 3)

protein depends on experiment on gel (Lilley et al. 2002). However, detection limit for silver staining is 1 ng. In saturation labeling cysteine residues of protein samples react with thiol reactive maleimide group of CyDye. CyDye supplied for saturation labeling does not contain any charge, thus binding does not affect the pI of protein samples. This is much more sensitive than minimal labeling because more fluorophor is incorporated in each protein sample (Shaw et al. 2003).

There are specific excitation wavelengths for different CyDye (cy2, cy3 and cy5). Gel scanning at specific wavelength provides the information about different proteomes (Fig. 11.4). Softwares developed for the DIGE system are typically used for the analysis of gel image. Images analyzed by the software systems are overlaid and the difference in protein spots can be detected. After completion of electrophoresis, image is analyzed three times at three different wavelengths (red, blue, and green). Various free tools are available on internet to analyze and compare spots on gel. Some of the important tools are: (1) Flicker (open2dprot.sourceforge.net/Flicker) (2) Image master 2D platinum (www.gelifesciences.com) (3) Melanie viewer (www.expasy.org) (4) GelScape (www.gelscape.ualberta.ca) and (5) PDQuest (www.bio-rad.com). This labeling system is also compatible to MS/Tandem MS, which involves generation of peptides by proteolytic digestion on gel plug itself. Cleavage by trypsin takes place on carboxy terminal of lysine and arginine residues but peptide generation is not affected because very few lysine residues undergo modification by dye labeling. The peptides generated after proteolytic digestion are extracted from the gel and results are analyzed by MS/Tandem MS. These results are compared *in silico* to identify the protein of interest (Fig. 11.5).

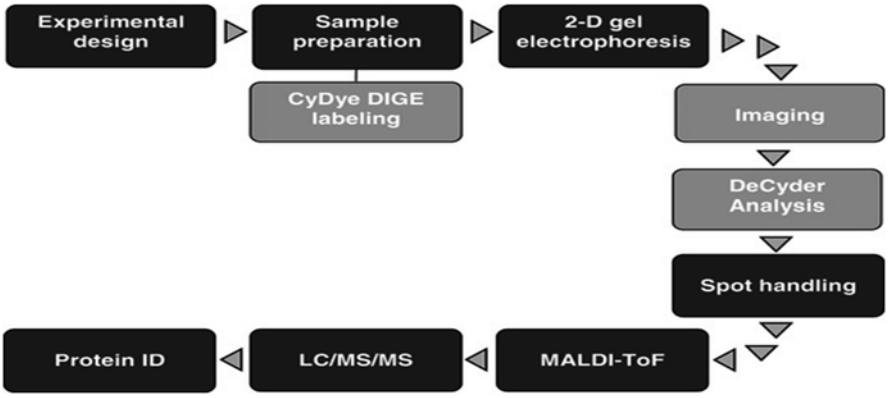

Fig. 11.5 Overview of the proteomics 2D workflow and 2D DIGE system approach for differential analysis and identification of protein samples. (Figure and figure legend are adopted from *Analytical and Bioanalytical Chemistry*, 2005, 382)

There are two important methods to identify proteins: (1) MALDI-Tof based peptide mass fingerprinting and (2) LC-MS/MS based peptide sequencing. Due to high resolution, sensitivity, and mass accuracy of MALDI-Tof, still this is one of the preferred methods to identify protein and also known as peptide mass fingerprinting. In this method, first protein of interest is either chemically or enzymatically digested and one MS-Spectrum is acquired which generates the mass per unit charge (m/z) of all peptide fragments (Fig. 11.6). These experimentally generated m/z of peptide fragments are identified by comparing to the list of identified peptide masses available in databases.

In LC-MS/MS, peptide fragments generated after proteolytic digestion are separated on HPLC system and eluted peptide from HPLC is fragmented by tandem mass spectrometry attached to HPLC by a process called collision-induced-dissociation (Fig. 11.7). MS/MS spectra are generated for individual peptides are compared to the list of peptides in databases. In addition to peptide mass, the peak pattern in CID spectra provides useful information about the sequence of protein.

In PMF analysis, the experimental result is compared to theoretical one that is stored in protein database using *in silico* digestion of same proteolytic enzyme used in experimental digestion. Some of the important tools available online for the PMF analysis are: (1) MS-Fit (prospector.ucsf.edu) (2) Mascot (www.matrixscience.com) (3) PeptideSearch (4) ProFound (prowl.rockefeller.edu) (5) Aldente (6) PepFrag (prowl.rockefeller.edu/prowl/pepfrag.html). This analysis considers the overlapping masses among the experimental and theoretical peaks and provides a similarity score of the proteins. There are variety of different scoring types available based on various algorithms used in above mentioned software. PepFrag and ProFound are based on simple scoring algorithm based on overlapping masses between experimental and theoretical digested protein whereas Mascot, MS-Fit, and ProFound are based on different algorithm, and it considers nonuniform distribution of peptide masses in databases. Aldente uses mass spectral realignment

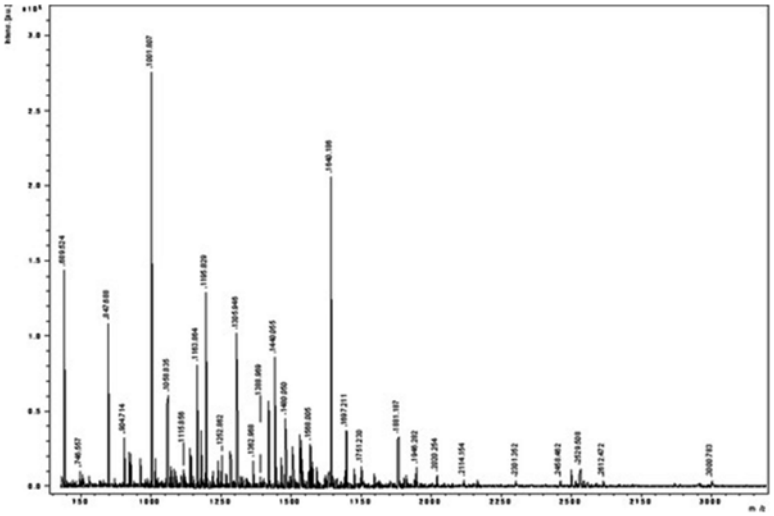

Fig. 11.6 MALDI-TOF-MS spectra of protein in gel tryptic digestion. (Figure and figure legend are adopted from *Applied Biochemistry and Biotechnology*, 2012, 167)

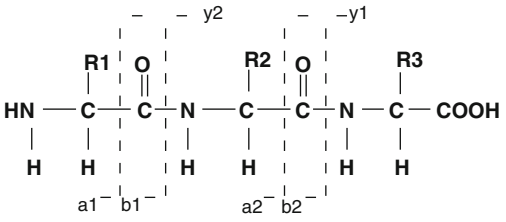

Fig. 11.7 Typical ion fragments observed for low-energy, collision-induced ion dissociation. (Figure and figure legend adopted from *Methods in Molecular Biology*, 2010, 593)

method also called Hough transformation in the scoring process (Gras and Muller 2001). Swiss-prot group allows Aldente to automatically search for post-translational modification in proteins as available in swiss-prot databases. It has been reported that Profound is much more sensitive and specific than MS-Fit for PMF analysis (Chamrad et al. 2004). In LC-MS/MS based peptide sequencing also called as peptide fragment fingerprinting (PFF), experimental MS/MS spectra are compared to theoretical MS/MS spectra of peptides from databases in order to find the similar peptide and get the highest score. Identification of protein based on fragmentation spectra has several advantages over peptide mass fingerprinting. PFF allows to work with complex peptide mixtures and to search local homology in databases. Due to the presence of contaminants, co-eluting peptides, post-translational modifications, mutations, and experimental spectra that are not available in databases, the theoretical MS/MS spectra do not match with experimental MS/MS spectra. Important tools available for PFF analysis are: (1) MS-Tag and MS-Seq (prospector.ucsf.edu) (2) Mascot (www.matrixscience.com/search_form_select.html) (3) GutenTag (fields.scripps.edu) (4) PepFrag (prowl.rockefeller.edu/prowl/pepfrag.html) (5) Sequest (fields.scripps.edu/sequest/index.html) (6) InsPecT (proteomics.ucsd.edu/

Software/Inspect.html) (7) Phenyx (www.phenyx-ms.com) (8) Popitam (www.code.google.com/p/popitam) (9) ProID (www.sashimi.sourceforge.net/software_mi.html) (10) X!Tandem (human.thegpm.org/tandem/thegpm_tandem.html) (11) Spectrum Mill (www.home.agilent.com).

The function of all these tools is based on database digestion, overlapping peptide sequences, scoring and validation of result. Guten Tag, InsPect, and Popitam are designed in such a way to handle unexpected post-translational modifications and mutations whereas Mascot and X!Tandem first identify the matched spectra and then it starts comparing the unidentified spectra using various parameters considering certain modifications as well as mutations. In the abovementioned tools, Sequest and Spectrum Mill have very good sensitivity values, whereas Mascot and X!Tandem have very good specificity values. Similar MS based tools generate relatively inconsistent results and become highly laborious to analyze the data. Recent advancement in tools such as Protein Prophet and Peptide Prophet has been built to help in validating their results. There are numerous tools available for MS data analysis. These tools are: (1) PeptideProphet (tools.proteomecenter.org/PeptideProphet.php) (2) ProteinProphet (tools.proteomecenter.org/ProteinProphet.php), and (3) DTASelect (fields.scripps.edu), for validation of protein identifications, while (4) RelEx (fields.scripps.edu) (5) ZoomQuant (Proteomics.mcw.edu) and (6) Xpress (tools.proteomecenter.org/XPRESS.php) for quantitative analysis of proteins obtained by MS spectra. ProteinScape (www.bruker.com/products/mass-spectrometry-and-separations/software/proteinscape/overview.html) is a commercial bioinformatics tool from Bruker Daltonics that manages data storage, processing, and identification of MS data. It combines the results obtained from various identification tools and generates a combined unique score and unidentified spectra that can be interpreted using algorithm. Scaffold 4 (www.proteomesoftware.com/products/scaffold), another bioinformatics tool from Proteome Software that identifies regulated isoforms and post-translational modifications of proteins, analyzes Mascot and Sequest results and filters by selecting target peptide by using new local FDR-based peptide scoring algorithm.

Bioinformatics Tool to Analyze Genomics Data

The reward of being the basic structural and functional unit of living organisms is given to “cell.” Each cell is an amazing entity in itself as it carries different organelles which interact and co-ordinate to ensure the ultimate cell survival. All cell functions are governed by the set of genetic instructions originating from DNA and ultimately moving to active proteins via RNA intermediary. These genetic materials (DNA & RNA) of an organism, instructing cells for what to do and how to do, constitute the genome of that organism. The study of genome is termed “Genomics,” which started way back in 1970s when Fred Sanger sequenced the genomes of a virus and mitochondrion. 1990 marks the initiation of human genome project which has generated large volume of data on genomic sequences. Since then the novel sequence submission has increased enormously which is widening the gap between

the volume of raw data and their analysis using techniques such as molecular biology and other traditional research approaches. This challenging and daunting interpretation of large amount of gene sequences can be achieved rapidly by using bioinformatics methodologies which can dissect large gene list to summarize most pertinent and enriched biology. There are a variety of genomic databases, gene analysis, and enrichment technologies which have become excellent tools to answer many genome related queries.

Genomic Repositories

The raw sequence data generated through genomic research are stored in genomic repositories. These public databases are classified on the basis of those containing primary data and those housing compilation of more curated version of data. The most common databases used for storing primary data are National Centre for Biotechnology Information (NCBI), European Molecular Biology Laboratory (EMBL), and DNA Database of Japan. These three repositories work together by sharing information on daily basis which is achieved by making files and sequence information compatible between individual databases. These databases monitor submission from researchers, genome projects, and other sources while doing little with curation part. The institutions such as NCBI are working to create more curated version of databases such as RefSeq, etc. RefSeq contains non-redundant, well-annotated, and curated collection of sequences including DNA, RNA, and their protein products.

Similarity Search and Sequence Alignment Tools

Basic local alignment search tool (BLAST) and its derivatives such as ClustalW, FASTA, Gapped BLAST, etc. are the most commonly used algorithms for alignment of sequences and similarity search. BLAST uses heuristic approach which accounts for a balance between speed and sensitivity. This heuristic approach uses short segments to create alignment instead of comparing every residue against each other. A word list with words of a specific length is created from the query sequence. The words from this word list are compared in order to seed an alignment. Once the alignment is seeded, an optimal final alignment is generated by further extending each match. Quality of the alignment is determined by both e- value and bit score. Bit score is an indicator of quality of an alignment; the higher the bit score, the better the alignment. The *e*-value of a given pair wise alignment is a measure of its statistical significance where the lower *e*-value indicates that the hit is more significant. An *e*-value of 0.05 indicates that this match has 5 in 100 chance of occurring by chance alone. Similarly FASTA also uses heuristic approach for sequence alignment. Gapped BLAST allows insertions and deletions within an alignment (Teufel et al. 2006). ClustalW is a multiple sequence alignment tool which helps in aligning

three or more biological sequences of comparable length. This helps in inferring homology and evolutionary relationships between query sequences.

The alignment of proteins or cDNA derived sequences with respect to genomic DNA sequences and those of nucleotide sequences against proteins is of increasing importance. EST_GENOME and SIM4 are the software tools which align transcribed and spliced DNA sequences to unspliced genomic DNA sequences containing that gene thereby allowing incorporation of introns where feasible intron starts and stops at splice consensus dinucleotides GT and AG, respectively. Other more developed and elaborated software for aligning genomic sequences with proteins are GeneWise and Procrustes (Searls 2000).

Variation Related Databases

Mutation is associated with various pathological conditions and hence seeks great importance to be identified and analyzed. Single nucleotide polymorphisms (SNPs) are also very important in relation to their pharmacological impact and role in disease development and drug interaction. One of the most comprehensive repositories for SNPs is NCBI dbSNP database. SNPs can be functionally silent if they are present in introns or in coding sequence positions where they do not account for change in amino acid in translated proteins. However, functionally active SNPs are present in regulatory regions of genes or in amino acid coding positions. There are other databases containing information regarding variety of known gene lesions. The human gene mutation database (HGMD) contains a compilation of data on germline mutation underlying human inherited diseases. It comprises known single base-pair substitutions, deletions, insertions, repeat expansions, etc., excluding mitochondrial genome mutations and somatic gene mutations. Each entry in HGMD comprises a reference to the first report of that mutation, the associated pathological state, chromosomal location, the gene name, and symbol (Cooper et al. 1998; Stenson et al. 2009). Another commonly used database for variation analysis is Online Mendelian Inheritance In Man (OMIM) which contains summary of genes present in human and their related genetic disorders. The OMIM entries are classified on the basis that the information they contain is related to genes, phenotypes or both. Each OMIM entry is given a unique six digit number where the first digit is indicative of the type of inheritence i.e X-linked, Y-linked, autosomal, or mitochondrial (Hamosh et al. 2005; Amberger et al. 2009). Other important mutation databases are Japanese single nucleotide polymorphism (JSNP) database and HGVBase.

Gene Prediction Tools

Gene prediction is identifying coding regions of novel genes within genomic DNA. Gene prediction can be done in three ways: homology based, *ab initio*, and combination of both. *Ab initio* gene prediction can be achieved by a variety of

software such as GRAIL, GENESCAN, AUGUSTUS, etc. Gene Recognition and Analysis Internet Link (GRAIL) is a statistical measure of coding potential which has limitation of ignoring important biological knowledge of gene structure (Xu et al. 1996). GENESCAN and AUGUSTUS rely on Hidden Markov Model for gene prediction. Both these databases can be accessed through WWW (World Wide Web). These software use web interface to take up large genomic fragments and give back the predicted structures of exon-intron and promoters for both strands in the form of graphical output. It also generates a table which shows the positions of exon, intron, and promoter present within the sequence. Accuracy of these programs is excellent with individual gene, but it drops when genes are embedded in a lengthy genomic context. Homology based gene finding can be done by tools such as BLASTX, Procrustes, and GeneWise. BLASTX works by searching novel DNA sequences for introns similar to a known protein. These homology based tools work well with genes embedded with larger continuous genomic sequences. A combination of best of *ab initio* and homology based approach could form an ideal algorithm for gene prediction (Searls 2000; Teufel et al. 2006).

Expression Profiling Tools

Gene expression determination is an important parameter in comparing its importance in diseased and normal condition and hence exploring the potential of a particular gene as a target for development of future drugs. The level of gene expression depends on a variety of factors such as epigenetic modification, somatic and genetic variation, levels of specific transcription factors, etc., which limit its ab initio determination and analysis. Accurate examination of few genes has been achieved via certain algorithms such as Serial Analysis of Gene Expression (SAGE), high-throughput EST sequencing, gene expression microarray analysis, etc. The outputs generated from these analyses are stored in database repositories. SAGE uses digital analysis to analyze overall expression pattern of gene. SAGE methodology uses short sequence tags (10–14 bp) derived from 3′ end of m-RNA. These sequence tags are linked together to form a long serial molecule which is then cloned and sequenced. Using these tags, an allocation to individual gene can be done via database sequence alignment thereby allowing analysis of gene expression by quantitating the number of sequenced tags present in a gene. The data generated through SAGE and EST experiments can be accessed through different web interfaces such as NCI CGAP suite, Stanford SOURCE web tool, etc. NCI CGAP is a sequence oriented suite providing user access to each EST sequences with extra information such as vectors and RNA library used in the experiments. SOURCE web tool is a very convenient interface which allows searching of data on SAGE and EST expression using names and accession numbers of genes (Diehn et al. 2003 Teufel et al. 2006). ArrayExpress is a public database at EBI consisting of microarray gene expression data (MGED). ArrayExpress has major three objectives: (1) to act as repository of data to support scientific publications; (2) to provide easy access to good quality data on gene expression; (3) to enable sharing of microarray

experimental protocols and models. ArrayExpress embraces three kinds of submissions containing arrays, protocols, and experiments each of which can be submitted individually and is given a unique identifier. ArrayExpress implements two standards: The Minimum Information about a Microarray Experiment (MIAME) which is a standard for microarray data annotation and Microarray Gene Expression Markup Language (MAGE-ML) which is a data exchange format based on XML and is developed by MGED society and Object Management Group (OMG) (Brazma et al. 2003). Another tool developed for viewing and examining data on gene expression with respect to biological pathways is Gene Microarray Pathway Profiler (GenMAPP). In this case, the gene expression data on pathways is displayed by color coding based on the criteria and data provided by the investigators. It also contains graphic tools for creating and modifying pathways (Dahlquist et al. 2002).

Tools for Promoter Prediction

The prediction of promoter region is very important as promoter is essential for the regulation of gene expression. Recognition of promoter is a tedious and problematic task owing to the following problems:

1. The position of promoter cannot be specified unless the gene is well characterized as the 5′ end of mRNA which is the transcription initiation point may not be known with accuracy.
2. The sequence is not very closely related to a known consensus sequence.
3. The promoter sequence can be considerable distance, may be kilobases, upstream of coding sequence of gene which is attributed to the possibility of splicing within 5′ leader and anywhere beyond the initiation point.
4. Sequence of mammalian promoters do not show similarity.

To solve these structural problems, most bioinformatics tools focus on core regions of promoters. Few important structural regions of promoter are TATA box and transcription factors binding sites. TATA binding proteins (TBP) recognizes TATA box and is a part of TFIID transcription initiation factor. It makes up the TATA box consensus sequence which can be used as an identifying feature of promoter. An initiator region which is a loosely conserved region around the transcription start site has also been reported which may be bound to different proteins and help in the recognition of transcription start site in absence of TATA box. The presence of promoters can also be signaled by presence of specific binding sites for transcription factors in vicinity of coding sequences. The common core promoter elements are represented in Fig. 11.8.

The computational approach for promoter recognition combines modules recognizing every binding site by using complete description of spatial arrangement of these binding sites. The binding specificity is sometimes characterized by using consensus sequence, that is, by assigning most favored base at a particular position within a site. Another way to determine binding specificity is by using position weight matrix. A position weight matrix assigns a weight to every feasible nucleo-

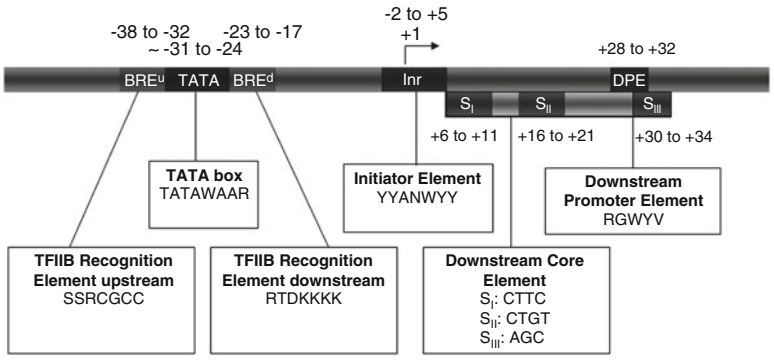

Fig. 11.8 Common core promoter elements with their respective consensus sequences and their relative positions with respect to transcription start site (+1). BREu and BREd bind with TFIIB transcription factor whereas the TATA box, Inr, DPE, and DCE are recognition site for TFIID. (Figure adopted from Molecular Biotechnology, 2010, 45:3)

tide at each position of potential binding site and the summation of these weights given to nucleotides are allotted as a site score. Position weight matrix is known to be the more informative description of specificity protein's binding to DNA than is a consensus sequence. The first and highly specific promoter prediction algorithm was PromoterScan. It recognizes promoters by means of TATA box position weight matrix in combination with the density of specific transcription factor binding sites. PromoterScan is considered to show high specificity but low sensitivity. Autogene is software including a module for promoter recognition. It uses a set of 136 consensus sequences for transcription factor binding sites assembled by Faisst and Meyer in 1992. PromFind works by using the differences in hexamer sequence frequencies between promoters, coding regions, and noncoding regions present downstream of the first exon (Fickett and Hatzigeorgiou 1997).

Genome Annotation Tools

The high rate of raw sequence accumulation necessitates the need of fully automated genome annotation to minimize the manual intervention. On-line annotation, also called as framework annotation can be achieved by using tools such as PowerBLAST. It breaks down a long query sequence into overlapping fragments, uses Gapped BLAST for searching, and then reassembles the results. Its search is based on taxonomical information and offers graphical display where annotations are superimposed on sequences. It also has choice for masking repetitive and low complexity sequences. AceDB is a dedicated on-line annotation system designed to support the *C. elegans* project. It depends on *ab initio* and hybrid gene finding tools with an option to use detectors for certain auxiliary feature such as tRNA finders. An attempt to annotate human genomic sequence was made by establishing a consortium known as Genome Channel. Certain tools such as GeneQuiz and MIPS are

Fig. 11.9 The tool menu page of DAVID showing that the gene list manager is displaying various tools after gene lists are successfully submitted. (Figure adopted from *Methods in Molecular Biology*, 2012, 820)

used as post hoc annotation systems. For a set of open reading frame, these tools provide functional predictions involving function, catalytic activity, cofactors, active sites, homology to other proteins, quaternary structures, etc. (Searls 2000). In order to ease the analysis of large list of genes and facilitate their functional annotation, a Database for Annotation, Visualization, and Integrated Discovery (DAVID) has been developed. It offers a set of data mining tools that provides a graphical display of functionally descriptive data. DAVID provides visualization tools which is linked to different sources of biological annotation and promotes discovery through functional classification, conserved protein domain architecture, and biochemical pathway maps (Dennis et al. 2003). It accelerates the functional annotation of any set of genes from human, mouse, and rat or flies genomes (Huang et al. 2009). The tool menu page of DAVID is shown in Fig. 11.9.

Acknowledgement PhD fellowship to RK and SS by Indian Institute of Technology Guwahati is acknowledged. RK and SS wrote the manuscript. VKD edited the draft.

References

Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance In Man (OMIM®). Nucleic Acids Res 37(Database Issue):D793–D796

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res 33(Database Issue):D562–D566

Baumann M, Pontiller J, Ernst W (2010) Structure and basal transcription complex of RNA polymerase II core promoters in the mammalian genome: an overview. Mol Biotechnol 45(3):241–247

Boonmee A, Srisomsap C, Chokchaichamnankit D, Karnchanatat A, Sangvanich P (2011) A proteomic analysis of Curcuma comosa Roxb rhizomes. Proteome Sci 9(43):1–8

Brazma A, Parkinson H, Sarkans U et al (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31(1):68–71

Chamrad DC, Korting G, Stuhler K et al (2004) Evaluation of algorithms for protein identification from sequence databases using mass spectrometry. Data Proteomics 4:619–628

Chang DW, Colucci G, Vaisar T et al (2007) Proteomic analysis of two non-bronchoscopic methods of sampling the lungs of patients with the acute respiratory distress syndrome (ARDS). Clin Proteomics 3:30–41

Cooper DN, Ball EV, Krawczak M (1998) The human gene mutation database. Nucleic Acids Res 26(1):285–287

Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 31:19–20

Dennis G, Sherman BT, Hosack DA (2003) DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 4(9):R60.1–R60.11

Diehn M, Sherlock G, Binkley G et al (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. Nucleic Acids Res 31(1):219–223

Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. Genome Res 7:861–878

Gras R, Muller M (2001) Computational aspects of protein identification by mass spectrometry. Curr Opin Mol Ther 3:526–532

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance In Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(Database Issue):D514–D517

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1):1–13

Imamichi T, Yang J, Huwang DW, Sherman B, Lempicki RA (2012) Interleukin-27 induces interferon-inducible genes: analysis of gene expression profiles using Affymetrix microarray and DAVID. Met Mol Biol 820:25–53

Kondo T, Hirohashi S (2009) Application of 2D-DIGE in cancer proteomics toward personalized medicine. Methods Mol Biol 577:135–154

Lilley KS, Razzaq A, Dupree P (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation. Curr Opin Chem Biol 6:46–50

Marouga R, David S, Hawkins E (2005) The development of the DIGE system: 2D fluorescence difference gel analysis technology. Anal Bioanal Chem 382:669–678

Matthiesen R, Amorim A (2010) Proteomics facing the combinatorial problem. Methods Mol Biol 593:175–186

Morisawa H, Hirota M, Toda T (2006) Development of an open source laboratory information management system for 2-D gel electrophoresis-based proteomics workflow. BMC Bioinformatics 7(430):1–11

Newburger DE, Bulyk LM (2009) UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic Acids Res 37(Database Issue):D77–D82

Parkinson H, Sarkans U, Shojatalab M et al (2005) ArrayExpress – a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33(Database Issue):D553–D555

Raghavendra R, Neelagund SE (2012) Biochemical characterization of novel bioactive protein from silkworm (*Bombyx mori* L.) fecal matter. Appl Biochem Biotechnol 167:1002–1014

Reynolds JA, Tanford C (1970) The gross conformation of protein-sodium dodecyl sulfate complexes. J Biol Chem 245(19):5161–5165

Searls DB (2000) Bioinformatics tools for whole genome. Annu Rev Genomics Hum Genet 1:251–279

Shaw J, Rowlinson R, Nickson J et al (2003) Evaluation of saturation labeling two dimensional difference gel electrophoresis fluorescent dyes. Proteomics 3:1181–1195

Shinde K, Phatak M, Johannes FM et al (2010) Genomics portals: integrative web-platform for mining genomics data. BMC Genomics 11:27

Stenson PD, Mort M, Ball EV et al (2009) The human gene mutation database: 2008 update. Genome Med 1(1):13.1–13.6

Teufel A, Krupp M, Weinmann A, Galle PR (2006) Current bioinformatics tools in genomic biomedical research (Review). Int J Mol Med 17:967–973

Xu Y, Mural RJ, Einstein R, Shah MB, Uberbacher EC (1996) GRAIL: a multi-agent neural network system for gene identification. Proc IEEE 84(10):1544–1552

Chapter 12
High-Throughput Transcriptome Analysis of Plant Stress Responses

Güzin Tombuloğlu and Hüseyin Tombuloğlu

Plant Stresses and Its Genetic Regulation

Plants are often subjected to various biotic and abiotic stresses in their natural or agronomic habitats (Ahuja et al. 2010; Rasmussen et al. 2013). Unfavorable environmental conditions cause serious limitations to agricultural production and crop yield (Cabello et al. 2014). Some examples of abiotic stresses are drought, salinity, extreme temperatures, chemical toxicity, and oxidative stress that have critical threats to agriculture and ruin the environment (Wang et al. 2003).

Abiotic stress that plays a major role in crop loss worldwide, reducing productivity of yields for most important crop plants by more than 50 % (Boyer 1982; Bray et al. 2000). Especially drought and salinity are assumed to become widespread in several regions. Nearly 22 % of the agricultural area is saline and also drought lands are expected to increase (FAO Food, Agriculture Organization of the United Nations 2004). And by the year 2050 it is assumed that more than 50 % of all plowlands will be subjected to salinity problem (Bray et al. 2000). Abiotic stress causes to several morphological, physiological, biochemical, and molecular changes (Wang et al. 2001).

The types of abiotic factors such as drought, salinity, cold, heat, and chemical pollution are often interactively contributed to similar cellular damage. This situation continues with the emergence of osmotic stress and in this concept homeostasis and ion concentration of cell is destroyed (Serrano et al. 1999; Zhu 2001a). Oxidative stress emerges to response of abiotic stress factors and may induce denaturation

G. Tombuloğlu, Ph.D.
Vocational School of Medical Sciences, Fatih University, Istanbul, Turkey

H. Tombuloğlu, Ph.D. (✉)
Department of Biology, Fatih University, 34500 Büyükçekmece, Istanbul, Turkey
e-mail: gkekec@fatih.edu.tr

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_12

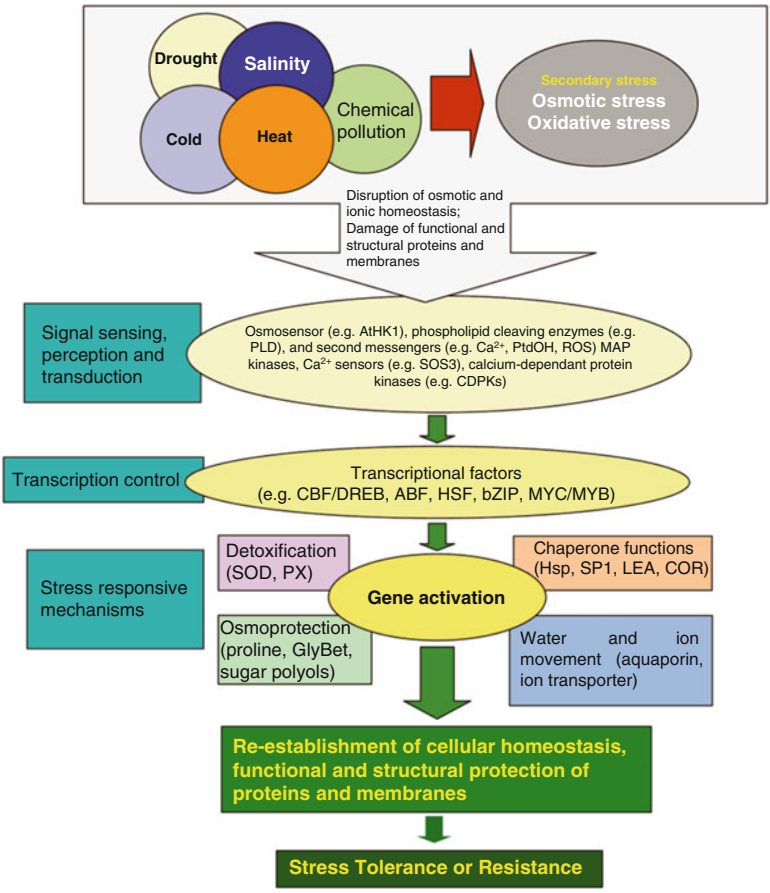

Fig. 12.1 Plant responses to abiotic stress (Wang et al. 2003)

of functional and structural proteins (Smirnoff 1998). Eventually different stress conditions often induce similar signaling pathways (Shinozaki and Yamaguchi-Shinozaki 2000; Knight and Knight 2001; Zhu 2001b, 2002) and cellular responses, such as production of stress proteins, antioxidants, and compatible solutes (Wang et al. 2003).

Genetic regulation of plants to abiotic stress is controlled by the activation of multiple responses involving complex gene interaction and crosstalk with many molecular pathways to increase the tolerance in order to adapt to adverse conditions (Basu 2012; Umezawa et al. 2006). Many genes and biochemical-molecular mechanisms are induced in plant response to abiotic stress are illustrated in Fig. 12.1. The primary stress signals such as changes in osmotic and ionic strength and membrane fluidity evolve to signal sensing, perception, and transduction (Wang et al. 2003).

Engineering of more tolerant plants involve illumination of specific stress-related genes related to molecular control mechanism of abiotic stress tolerance. These stress-related genes can be classified in three main groups: The first group is activated in signaling cascades and in transcriptional control, such as MyC, MAP kinases and SOS kinase (Shinozaki and Yamaguchi-Shinozaki 2000; Munnik et al. 1999; Zhu 2001b), phospholipases (Chapman 1998; Frank et al. 2000), and involving several transcriptional factors such as HSF, and the CBF/DREB and ABF/ABAE families (Stockinger et al. 1997; Schoffl et al. 1998; Shinozaki and Yamaguchi-Shinozaki 2000). The second group is directly perform protection of membranes and proteins, such as heat shock proteins (Hsps) and chaperones, late embryogenesis abundant (LEA) proteins (Bray et al. 2000), osmoprotectants, and free-radical scavengers (Bohnert and Sheveleva 1998). Final groups are involved in water and ion uptake and transport such as aquaporins and ion transporters (Wang et al. 2003). In this concept hormonal responses and reactive oxygen species (ROS), which is a powerful signaling molecule, reveal regulatory networks. Finally stress-responsive mechanism is activated to re-establish cellular homeostasis by the regulatory components of stress-induced genes. And also this mechanism excretes toxic components and protects and repairs damaged proteins and membranes (Cabello et al. 2014).

Transcriptome Analysis Upon Stress Conditions

Transcriptome discovery provides deep insight to understanding stress response of an organism (Jogaiah et al. 2013). Transcriptome analysis is a powerful functional genomics technology has been used to identify plant stress adaptation and tolerance mechanisms in recent studies (Urano et al. 2010). Mizuno et al. (2010) performed whole mRNA sequencing and identified salinity stress-inducible genes in rice on the basis of piling up of mapped reads. By this study, it was contributed to discover unannotated transcripts which had ORFs with a mean length of 123 amino acids in root and 125 amino acids in shoot could encode functional proteins without prior annotations. 213 and 436 unannotated transcripts were differentially expressed in shoot and root tissues, respectively. This comprehensive study demonstrated that the accuracy of mRNA sequencing technology in differentially expressed genes without a limitation in analysis annotated genes by the comparison of array-based and mRNA sequencing technology.

Dong et al. (2012) analyzed mRNA expression levels of *Sinapis alba* leaves under rewatering growth conditions and drought stress conditions by using high-throughput sequencing technology. It was reported that 557 annotated genes were involved in drought stress response related to signaling components, transcription regulators, or other proteins which were required for cell growth and development. This study provided candidate genes for understanding the molecular mechanisms of drought stress tolerance of *S. alba* by using high-throughput sequencing technology.

Zhang et al. (2012) investigated atrazine-responsive transcriptome in rice by high-throughput sequencing to utilize understanding of gene expression and regulatory mechanisms of plant adaptation to xenobiotic stress. The most differentially expressed 40 genes were demonstrated and encoding proteins or enzymes were indicated. Enrichment analysis of numerous genes was involved in metabolism of carbohydrates, organic acids, sulfate, amino acids, secondary metabolites, etc. These genes were categorized to response to intracellular and environmental stimulus, glutathione transferase activity, and oxidoreductase activity. Namely the high-throughput sequencing technology presents a major advantage for characterization of toxicological responses of crops to atrazine.

Newly, Xu et al. (2013) identified early response of highly tolerant *Gossypium aridum* to salt stress. Digital gene expression (DGE) analysis was performed to identify the genes involved in salt stress. Transport, response to hormone stimulus, and signaling were the distinctly indicated pathways under salt stress conditions. According to the GO analysis, protein kinase activity and transporter activity were the most highly enriched GO terms. This study was the first report of root and leaf transcriptome characterization of highly tolerant *G. aridum* by using paired-end sequencing technology.

Transcriptome De Novo Assembly

In recent years next generation sequencing (NGS) technology has become a magnificent approach in the genome and transcriptome analysis of any organism (Wang et al. 2010). The current NGS technologies are reference-based, de novo and combined strategies differ in throughput and cost property (Martin and Wang 2011).

Reference-based assembly is very sensitive and contains three steps when a reference genome is available (Martin and Wang 2011). First step is the alignment of RNA-seq reads by using a splice-aware aligner. Second step is building a graph representing alternative splicing. And final step is traversing the graph to assemble variants and getting assembled isoforms. Reference-based transcriptome assembly is useful for simple transcriptomics organisms such as bacteria, archaea, and simple-eukaryotic organisms related to lower introns and alternative splicing but it is difficult to use for the plant and mammalian transcriptomes related to complex alternative splicing (Martin and Wang 2011). However, this technique is only applicable when the reference genome is available (Trapnell et al. 2010).

De novo assembly of RNA-seq was discovered as a wonderful approach to high-throughput gene discovery on a genome-wide scale in non-model organisms. De novo assembly of RNA-seq is the only way to study transcriptomes of organisms without a reference genome (Grabherr et al. 2011). Recently, a number of de novo assemblers have been developed to give researchers an explicit understanding of process. These assemblers are Velvet (Zerbino and Birney 2008), Oases (Schulz et al. 2012), SOAPdenovo (Li et al. 2009), Rnnotator (Martin et al. 2010), and finally Trinity (Grabherr et al. 2011).

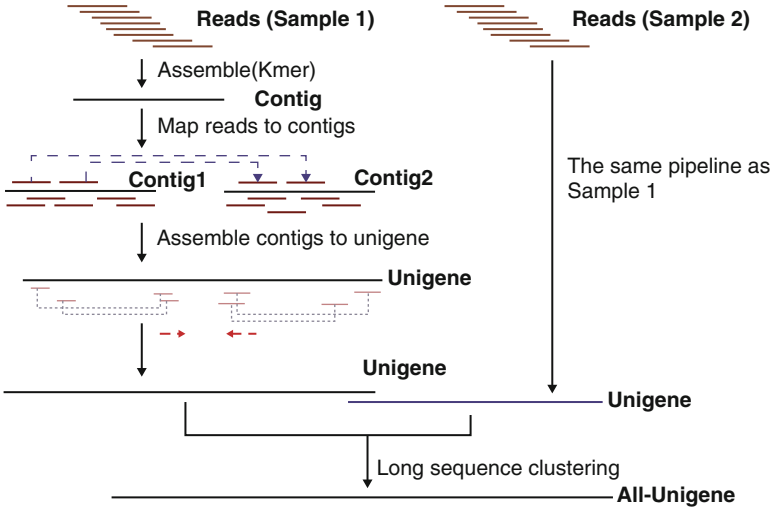

Fig. 12.2 Trinity assembly process (Tombuloglu 2014)

Trinity program is the efficient method in order to do full-length transcriptome assembly from RNA-Seq data without a reference genome (Grabherr et al. 2011). Firstly, from the RNA-Seq data, the program combines short reads to generate contigs which are longer fragments overlapped in certain lengths. Then the reads mapped back to contigs with paired-end reads. Finally, the program connects the contigs, and get sequences that corresponded to both contig ends. The corresponded sequence is called Unigene (Fig. 12.2). In the case of multiple sequence reading from the same species, Unigenes from different samples can be overlapped by using sequence clustering programs and then spliced and redundant sequences can be removed. Hence, non-redundant sequences can be acquired. After clustering, the non-redundant unigenes can be divided into two: one is clusters, referred as CL; and the other is singletons, called either Unigene or Pre. Finally, to detect the sequence directions, BLASTx alignment with e-value score <0.00001 is used to show the best aligning score between Unigenes and protein databases such as nr, Swiss-Prot, KEGG and COG. If there is a contradiction between different databases, the sequence direction can be decided by following the database order: nr, Swiss-Prot, KEGG, and COG. If a Unigene is not aligned with the databases, another software namely ESTScan (Iseli et al. 1999) can be used to decide its sequnce direction.

De novo assembly has been used to transcriptome sequencing in many organisms, such as chestnut (Barakat et al. 2009), pine (Parchman et al. 2010), olive (Alagna et al. 2009), ginseng (Sun et al. 2010), *Arabidopsis thaliana* (Weber et al. 2007; Wall et al. 2009), maize (Vega-Arreguin et al. 2009), *Artemisia annua* (Wang et al. 2009), fish (Hale et al. 2009), insects (Hahn et al. 2009; Zagrobelny et al. 2009), and worms (Meyer et al. 2009).

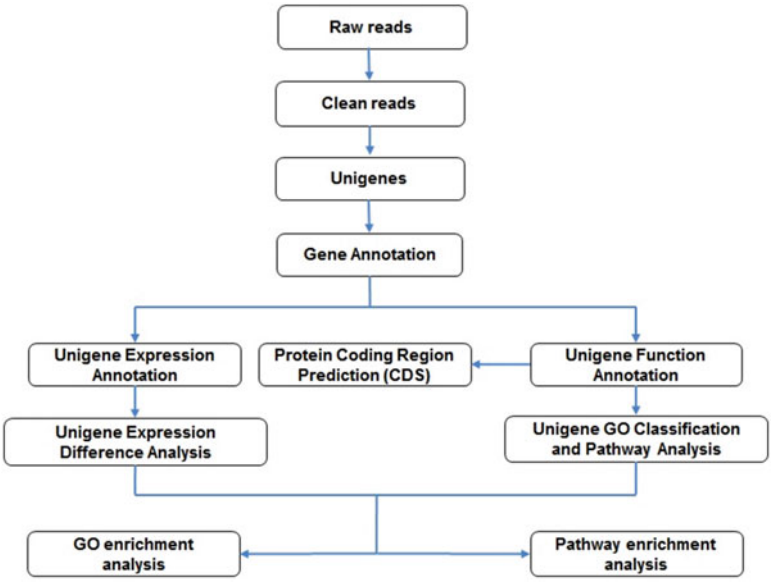

Fig. 12.3 Bioinformatics analysis (Tombuloglu 2014)

A. thaliana was the first plant in de novo plant transcriptome sequencing by mRNASeq (Weber et al. 2007). This novel approach used for molecular plant breeding studies such as for eucalyptus, melon, and different legumes (Novaes et al. 2008; Guo et al. 2010; Blavet et al. 2011; Hiremath et al. 2011; Kaur et al. 2011). Also this technology was applied to the primary (Dai et al. 2011; Franssen et al. 2011; King et al. 2011; Troncoso-Ponce et al. 2011) and the secondary (Alagna et al. 2009; Wang et al. 2009; Bleeker et al. 2011; Desgagne-Penix et al. 2012) metabolisms of plants. In a while, it has been used to study adaptations of plants to biotic (Barakat et al. 2009; Sun et al. 2011) and abiotic stress conditions (Dassanayake et al. 2009; Villar et al. 2011).

In this concept, we performed a transcriptome-based analysis of barley (*Hordeum vulgare*) exposed to excess Boron (B) (Tombuloglu 2014, Tombuloglu et al. 2015). Totally, 256,874 unigenes were generated and assigned to known peptide databases: Gene Ontology (GO), Swiss-Prot, Clusters of Orthologous Groups (COG), and the Kyoto Encyclopedia of Genes and Genomes (KEGG), as determined by BLASTx search (Fig. 12.3). DGE analysis was performed to identify the genes involved in excess B. Most of them were involved in the cell wall, stress response, membrane, protein kinase, and transporter mechanisms (Tombuloglu et al. 2013, 2015).

Since the first plant transcriptome, *Arabidopsis*, was sequenced (Weber et al. 2007), at least 60 plant transcriptomes have been sequenced by using de novo assembly. At this time, the purpose of the 1KP project is sequencing the transcriptome of about 1,000 plant species (Schliesky et al. 2012).

Functional Annotation of Unigenes by BLASTx Against Protein Databases

To annotate unigenes, all unigene sequences are firstly aligned by BLASTx to protein databases like nr, Swiss-Prot, KEGG, and COG (e-value < 0.00001), and aligned by BLASTn to nucleotide databases nt (e-value < 0.00001), retrieving proteins with the highest sequence similarity with the given Unigene annotations provide functional annotations of All-Unigene and expression levels (Fan et al. 2013). The nr nucleotide database is most widely used to search homologous proteins, preserved by NCBI as a target database for their BLAST search services from GenBank, GenBank updates, and EMBL updates (Lin et al. 2013). Using this approach, most researchers found the matches of unique sequences in the nr database in different organisms (Hao et al. 2011; Fan et al. 2013; Lin et al. 2013). After the alignment by BLASTn to nucleotide database processes, all unigenes were also searched against the Swiss-Prot database. Swiss-Prot database serves a high level of protein annotation by the integration of other databases. These annotations are functions of proteins, posttranslational modifications, domains and sites of proteins, secondary and quaternary structure of proteins, etc. (Bairoch and Apweiler 2000).

Functional Classification of Unigenes by Clusters of Orthologous Groups (COG), Gene Ontology (GO), and KEGG Pathway Enrichment

Unigene annotation provides information of expression and functions of Unigenes. Functional annotation of Unigene contains COG, Gene Ontology (GO), KEGG pathway enrichment, and protein sequence similarity.

COG is a database where orthologous gene products are classified. Every protein in COG is assumed to evolve from an ancestor protein, and the whole database is built on coding proteins with complete genome as well as system evolution relationships of bacteria, algae, and eukaryotic creatures (Tatusov et al. 1997, 2003). Unigenes are aligned to COG database to predict and classify possible functions of Unigenes.

Gene Ontology (GO) is a comprehensive database which classifies genes according their functions and properties in any organism (Ashburner et al. 2000). GO database is divided into three ontologies: molecular function, cellular component, and biological process. The database unit is defined as GO-term and each GO-term correspond to a type of ontology. For the annotation of All-Unigene, Blast2GO program is used with nr database (Conesa et al. 2005). After that, another program WEGO can be used to do GO functional classification for all Unigenes (Ye et al. 2006).

KEGG is a bioinformatics resource for linking genomes to life and the environment. KEGG database contains systematic analysis of inner-cell metabolic pathways and functions of gene products (Kanehisa et al. 2008). KEGG database enables to understand biological functions of genes with the Pathway annotation. The KEGG

pathway database records networks of molecular interactions in the cells, and variants of them specific to particular organisms (Kanehisa et al. 2010; Joy et al. 2013).

Protein Coding Region Prediction (CDS)

CDS represents the coding sequence of a protein (Furuno et al. 2003). It is composed of exons and the region starts from the 5′ end initiation codon and ends with 3′ end termination codon. Protein Coding Region Prediction is an important step in the analysis of functional gene annotation (Furuno et al. 2003). RNA-seq enables to obtain genomic information of an organism that is translated to protein from full-length mRNA (Grabherr et al. 2011).

In Protein Coding Region Prediction by using RNA-seq, all unigenes are firstly aligned by BLASTx to protein databases in the priority order of nr, Swiss-Prot, KEGG, and COG. Proteins with highest ranks in the blast results are taken to decide the coding region sequences of Unigenes. Amino acid translation of these sequences gives the coding region of a Unigene. When unigene cannot be aligned to any database, in that case the sequence can be scanned by ESTScan (Iseli et al. 1999) which gives nucleotide sequence (5′–3′) direction and amino acid sequence of predicted coding region.

Digital Gene Expression Profiling

Genome-wide gene expression analysis is the fundamental study in functional genomics research area. To determine expression level of the gene, microarrays are at present the default technology due to sequence-specific probe hybridization, measurement of the only relative abundances of transcripts and determination of only predefined sequences (Irizarry et al. 2005; Hoen et al. 2008).

Otherwise, tag-based methods including serial analysis of gene expression (SAGE) (Velculescu et al. 1995; Harbers and Carninci 2005) cap analysis of gene expression (CAGE) (Shiraki et al. 2003; Nakamura and Carninci 2004; Kodzius et al. 2006) and massively parallel signature sequencing (MPSS) (Brenner et al. 2000; Peiffer et al. 2008; Reinartz et al.; 2002) were generated to overcome these handicaps. These tag-based sequencing approaches are high throughput and can provide precise, “digital” gene expression levels (Wang et al. 2009). However, most of these methods are based on expensive Sanger sequencing technology. And it is impossible to map substantial part of the short tags uniquely to the reference genome (Wang et al. 2009). Also, only a part of the transcript can be analyzed. These properties are the limitations of the use of traditional sequencing technology (Wang et al. 2009).

By the improvement of high-throughput sequencing technology, Digital Gene Expression (DGE) profiling is started to use gene expression of whole genome in certain species under unique conditions. In DGE analysis specific tags representing

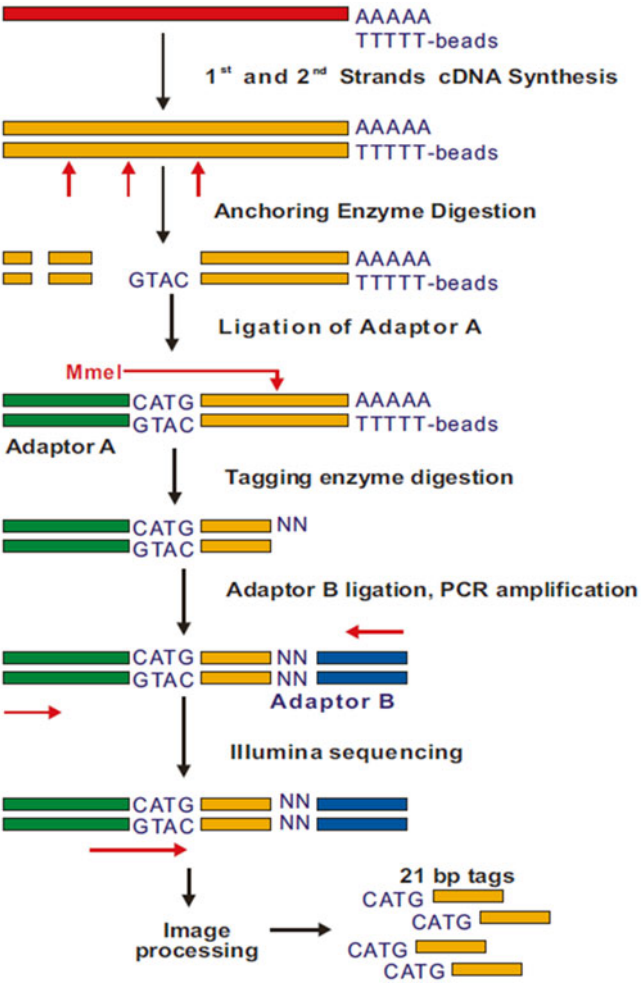

Fig. 12.4 Experimental workflow of DGE (BGI 2014, http://www.genomics.hk/Digital)

different mRNAs are acquired through efficient enzyme digestion. Then millions of tags were sequenced using the new-generation sequencing technology (Fig. 12.4).

The 3′ tag DGE uses oligo-dT priming for first strand cDNA synthesis generates libraries that are enriched in the 3′ untranslated regions of polyadenylated mRNAs, and produces 20–21 base cDNA tags (Saha et al. 2002). DGE technology generates such extensive sequencing depth-of-coverage that single copy resolution of gene expression quantification should be possible. The alternative to 3′ tag DGE sequencing is full-length RNA sequencing (RNA-seq), which collects both quantitative and qualitative information about the entire transcriptome. RNA-seq is a powerful tool for identifying expressed polymorphisms as well as differentially expressed splice

variants, fusion genes, transcriptional start sites, and polyadenylation sites (Wang et al. 2009; Asmann et al. 2009). With these advantages RNA-seq method provides both single-base resolution for annotation and "digital" gene expression levels at the genome scale, often at a much lower cost than either tiling arrays or large-scale Sanger EST sequencing (Wang et al. 2009).

Unigene Expression Difference Analysis

In all organisms every cell contains the same set of genes however only a certain part of these genes are activated in any given cell at a certain time. To identify which genes are turned on and which are turned off, expression difference analysis is used. By the discovery of differentially expressed genes that are in different conditions is a basic step of understanding the molecular basis of phenotypic variation (Soneson and Delorenzi 2013). For this purpose, the high-throughput sequencing of cDNA (RNA-seq) has been used extensively. The most common use of RNA-seq is finding differentially expressed genes with the quality of quantification that show differences in expression level between conditions related to response of any treatment (Oshlack et al. 2010; Agarwal et al. 2010; Soneson and Delorenzi 2013).

In expression difference analysis, normalization is the basic step in the analysis of expression difference from RNA-seq data and provides certain comparison of expression levels between and within samples (Sultan et al. 2008; Mortazavi et al. 2008; Marioni et al. 2008; Anders and Huber 2010; Langmead et al. 2010; Robinson and Oshlack 2010). Within-sample comparison library normalization enables quantification of relative expression levels of each gene to other genes in the sample. The RPKM (Reads Per kb per Million reads) measure of read density reflects the molar concentration of a transcript in the starting sample by normalizing for RNA length and for the total read number in the measurement. This facilitates transparent comparison of transcript levels both within and between samples (Mortazavi et al. 2008). The RPKM method is able to eliminate the influence of different gene length and sequencing level on the calculation of gene expression. Therefore the calculated gene expression can be directly used for comparing the difference of gene expression between samples. Thus the RPKM method allowed to study the expression levels of all the unigenes generated.

$$\text{The formula is RPKM} = (1{,}000{,}000 * C) / (N * L * 1{,}000)$$

Assigns RPKM(A) to be the expression of gene A, C to be number of reads that uniquely aligned to gene A, N to be total number of reads that uniquely aligned to all genes, and L to be number of bases on gene A.

Acknowledgment We would like to thank Prof. Arie Altman for the permission to use Fig. 12.1. Also, the authors acknowledge the financial support of the Scientific and Technological Research Council of Turkey (TUBITAK) (grant no: 111T015).

References

Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M (2010) Comparison and calibration of transcriptome data from RNA-seq and tiling arrays. BMC Genomics 11:383

Ahuja I, de Vos RC, Bones AM, Hall RD (2010) Plant molecular stress responses face climate change. Trends Plant Sci 15:664–674

Alagna F, Agostino ND, Torchia L, Servili M, Rao R, Pietrella M, Giuliano G et al (2009) Comparative 454 pyrosequencing of transcripts from two olive genotypes during fruit development. BMC Genomics 10:399

Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11:R106. doi:10.1186/gb-2010-11-10-r106

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Asmann YW, Klee EW, Thompson EA, Perez EA, Middha S et al (2009) 3′ tag digital gene expression profiling of human brain and universal reference RNA using Illumina genome analyzer. BMC Genomics 10:531. doi:10.1186/1471-2164-10-531

Bairoch A, Apweiler R (2000) The Swiss-Prot protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48

Barakat A, Diloreto DS, Zhang Y, Smith C, Baier K, Powell WA, Wheeler N et al (2009) Comparison of the transcriptomes of American chestnut (*Castanea dentata*) and Chinese chestnut (*Castanea mollissima*) in response to the chestnut blight infection. BMC Plant Biol 9:51

Basu U (2012) Identification of molecular processes underlying abiotic stress plants adaptation using "Omics" technologies. In: Benkeblia N (ed) Sustainable agriculture and new technologies. CRC, Boca Raton, pp 149–172

BGI 2014, http://www.genomics.hk/Digital.htm

Blavet N, Charif D, Oger-Desfeux C, Marais GAB, Widmer A (2011) Comparative high throughput transcriptome sequencing and development of SiESTa, the Silene EST annotation data base. BMC Genomics 12:376. doi:10.1186/1471-2164-12-376

Bleeker PM, Spyropoulou EA, Diergaarde PJ, Volpin H, De Both MTJ, Zerbe P, Bohlmann J, Falara V, Matsuba Y, Pichersky E, Haring MA, Schuurink RC (2011) RNA-seq discovery, functional characterization, and comparison of sesquiterpene synthases from *Solanum lycopersicum* and *Solanum habrochaites* trichomes. Plant Mol Biol 77:323–336

Bohnert HJ, Sheveleva E (1998) Plant stress adaptations—making metabolism move. Curr Opin Plant Biol 1:267–274

Bray EA, Bailey-Serres J, Weretilnyk E (2000) Responses to abiotic stresses. In: Gruissem W, Buchannan B, Jones R (eds) Biochemistry and molecular biology of plants. American Society of Plant Physiologists, Rockville, MD, pp 1158–1249

Brenner S et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634

Boyer JS (1982) Plant productivity and environment. Science 218:443–448

Cabello JV, Lodeyro AF, Zurbriggen MD (2014) Novel perspectives for the engineering of abiotic stress tolerance in plants. Curr Opin Biotechnol 26:62–70

Chapman D (1998) Phospholipase activity during plant growth and development and in response to environmental stress. Trends Plant Sci 3:419–426

Conesa A, Gotz S et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676

Dai N, Cohen S, Portnoy V, Tzuri G, Harel-Beja R, Pompan-Lotan M, Carmi N, Zhang GF, Diber A, Pollock S, Karchi H, Yeselson Y, Petreikov M, Shen S, Sahar U, Hovav R, Lewinsohn E, Tadmor Y, Granot D, Ophir R, Sherman A, Fei ZJ, Giovannoni J, Burger Y, Katzir N, Schaffer AA (2011) Metabolism of soluble sugars in developing melon fruit: a global transcriptional view of the metabolic transition to sucrose accumulation. Plant Mol Biol 76:1–18

Dassanayake M, Haas JS, Bohnert HJ, Cheeseman JM (2009) Shedding light on an extremophile lifestyle through transcriptomics. New Phytol 183:764–775

Desgagne-Penix MFK, Schriemer DC, Cram D, Nowak J, Facchini PJ (2012) Integration of deep transcriptome and proteome analyses reveals the components of alkaloid metabolism in opium poppy cell cultures. BMC Plant Biol 10:252

Dong CH, Li C, Yan XH, Huang SM, Huang JY, Wang LJ et al (2012) Gene expression profiling of *Sinapis alba* leaves under drought stress and rewatering growth conditions with Illumina deep sequencing. Mol Biol Rep 39:5851–5857

Fan H, Xiao Y, Yang Y, Xia W, Mason AS, Xia Z, Qiao F, Zhao S, Tang H (2013) RNA-seq analysis of *Cocos nucifera*: transcriptome sequencing and de novo assembly for subsequent functional genomics approaches. PLoS One 8:e59997

FAO (Food, Agriculture Organization of the United Nations) (2004) FAO production yearbook. FAO, Rome

Frank W, Munnik T, Kerkmann K, Salamini F, Bartels D (2000) Water deficit triggers phospholipase D activity in the resurrection plant Craterostigma plantagineum. Plant Cell 12:111–124

Franssen SU, Shrestha RP, Brautigam A, Bornberg-Bauer E, Weber APM (2011) Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. BMC Genomics 12:227. doi:10.1186/1471-2164-12-227

Furuno M, Kasukawa T, Saito R, Adachi J, Suzuki H, Baldarelli R, Hayashizaki Y, Okazaki Y (2003) CDS annotation in full-length cDNA sequence. Genome Res 13:1478–1487

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X et al (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat Biotechnol 29:644–652

Guo S, Zheng Y, Joung J, Liu S, Zhang Z, Crasta OR, Sobral BW et al (2010) Transcriptome sequencing and comparative analysis of cucumber flowers with different sex types. BMC Genomics 11:384

Hahn DA, Ragland GJ, Shoemaker DD, Denlinger DL (2009) Gene discovery using massively parallel pyrosequencing to develop ESTs for the flesh fly *Sarcophaga crassipalpis*. BMC Genomics 10:234

Hale MC, McCormick CR, Jackson JR, Dewoody JA (2009) Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. BMC Genomics 10:203

Hao DC, Ge GB, Xiao PG, Zhang YY, Yang L (2011) The first insight into the tissue specific taxus transcriptome via Illumina second generation sequencing. PLoS One 6:e21220. doi:10.1371/journal.pone.0021220

Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2:495–502

Hiremath PJ, Farmer A, Cannon SB, Woodward J, Kudapa H, Tuteja R, Kumar A, Bhanuprakash A, Mulaosmanovic B, Gujaria N, Krishnamurthy L, Gaur PM, Kavikishor PB, Shah T, Srinivasan R, Lohse M, Xiao YL, Town CD, Cook DR, May GD, Varshney RK (2011) Large-scale transcriptome analysis in chickpea (*Cicer arietinum* L.), an orphan legume crop of the semi-arid tropics of Asia and Africa. Plant Biotechnol J 9:922–931

Hoen PAC, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RHAM, de Menezes RX, Boer JM, van Ommen GJB, den Dunnen JT (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res 36(21):e141

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G et al (2005) Multiple-laboratory comparison of microarray platforms. Nat Methods 2:345–350

Iseli C, Jongeneel CV et al (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 1999:138–148

Jogaiah S, Govind SR, Tran LS (2013) System biology-based approaches towards understanding drought tolerance in food crops. Crit Rev Biotechnol 33:23–39

Joy N, Asha S, Mallika V, Soniya EV (2013) De novo transcriptome sequencing reveals a considerable bias in the incidence of simple sequence repeats towards the downstream of 'Pre-miRNAs' of Black Pepper. PLoS One 8(3):e56694. doi:10.1371/journal.pone.0056694

Kanehisa M, Araki M et al (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res 36(Database issue):D480–D484

Kanehisa M, Goto S, Furumichi M, Tanabe M and Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355–D360

Kaur S, Cogan NOI, Pembleton LW, Shinozuka M, Savin KW, Materne M, Forster JW (2011) Transcriptome sequencing of lentil based on second-generation technology permits large-scaleuni- gene assembly and SSR marker discovery. BMC Genomics 12:265. doi:10.1186/1471-2164-12-265

King AJ, Li Y, Graham IA (2011) Profiling the developing *Jatrophacurcas* L. seed transcriptome by pyrosequencing. Bioenergy Res 4:211–221. doi:10.1007/s12155- 011-9114-x

Knight H, Knight MR (2001) Abiotic stress signalling pathways: specificity and cross-talk. Trends Plant Sci 6:262–267

Kodzius R et al (2006) CAGE: cap analysis of gene expression. Nat Methods 3:211–222

Langmead B, Hansen KD, Leek JT (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. Genome Biol 11:R83. doi:10.1186/gb-2010-11-8-r83

Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967

Lin M, Lai D, Pang C, Fan S, Song M, Yu S (2013) Generation and analysis of a large-scale expressed sequence tag database from a full-length enriched cDNA library of developing leaves of *Gossypium hirsutum* L. PLoS One 8(10):e76443. doi:10.1371/journal.pone.0076443

Martin J, Bruno VM, Fang Z et al (2010) Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-seq reads. BMC Genomics 11:663

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Genetics 12:671–682

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517. doi:10.1101/gr.079558.108

Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics 10:219

Mizuno H et al (2010) Massive parallel sequencing of mRNA in identification of unannotated salinity stress-inducible transcripts in rice (*Oryza sativa* L.). BMC Genomics 11(683)

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods 5:621–628. doi:10.1038/nmeth.1226

Munnik T, Ligterink W, Meskiene I, Calderini O, Beyerly J, Musgrave A, Hirt H (1999) Distinct osmo-sensing protein kinase pathways are involved in signaling moderate and severe hyper-osmotic stress. Plant J 20:381–388

Nakamura M, Carninci P (2004) Cap analysis gene expression: CAGE. Tanpakushitsu Kakusan Koso 49:2688–2693 (in Japanese)

Novaes E, Drost DR, Farmerie WG, Pappas GJ Jr, Grattapaglia D, Sederoff RR, Kirst M (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312

Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. Genome Biol 11:220

Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11:180

Peiffer JA et al (2008) A spatial dissection of the *Arabidopsis* floral transcriptome by MPSS. BMC Plant Biol 8:43

Rasmussen S, Barah P, Suarez-Rodriguez MC, Bressendorff S, Friis P, Costantino P, Bones AM, Nielsen HB, Mundy J (2013) Transcriptome responses to combinations of stresses in *Arabidopsis*. Plant Physiol 161:1783–1794

Reinartz J et al (2002) Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. Brief Funct Genomic Proteomic 1:95–104

Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25. doi:10.1186/gb-2010-11-3-r25

Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20:508–512

Schliesky S, Gowik U, Weber APM, Bräutigam A (2012) RNASeq assembly—are we there yet? Front Plant Sci 3:220

Schöffl F, Prändl R, Reindl A (1998) Regulation of the heat-shock response. Plant Physiol 117:1135–1141

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092

Serrano R, Mulet JM, Rios G, Marquez JA, de Larrinoa IF, Leube MP, Mendizabal I, Pascual-Ahuir A, Proft M, Ros R, Montesinos C (1999) A glimpse of the mechanisms of ion homeostasis during salt stress. J Exp Bot 50:1023–1036

Shinozaki K, Yamaguchi-Shinozaki K (2000) Molecular responses to dehydration and low temperature: differences and cross-talk between two stress signalling pathways. Curr Opin Plant Biol 3:217–223

Shiraki T et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100: 15776–15781

Smirnoff N (1998) Plant resistance to environmental stress. Curr Opin Biotechnol 9:214–219

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91. doi:10.1186/1471-2105-14-91

Stockinger EJ, Gilmour SJ, Thomashow MF (1997) Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proc Natl Acad Sci USA 94: 1035–40

Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science 321:956–960. doi:10.1126/science.1160342

Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EM, Chen S (2010) De novo sequencing and analysis of the American ginseng root transcriptome using a GS FLX Titanium platform to discover putative genes involved in ginsenoside biosynthesis. BMC Genomics 11:262

Sun H, Paulin L, Alatalo E, Asiegbu FO (2011) Response of living tissues of *Pinus sylvestris* to the saprotrophic biocontrol fungus *Phlebiopsis gigantea*. Tree Physiol 31:438–451

Umezawa T, Fujita M, Fujita Y, Yamaguchi-Shinozaki K, Shinozaki K (2006) Engineering drought tolerance in plants: discovering and tailoring genes to unlock the future. Curr Opin Biotechnol 17(2):113–122

Urano K, Kurihara Y, Seki M, Shinozaki K (2010) 'Omics' analyses of regulatory networks in plant abiotic stress responses. Curr Opin Plant Biol 13:132–138

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

Tombuloglu H, Kekec G, Sakcali MS, Unver T (2013) Transcriptome-wide identification of R2R3-MYB transcription factors in barley with their boron responsive expression analysis. Mol Genetics Genomics 288:141-155

Tombuloglu G (2014) Transcriptomics identification of barley (Hordeum vulgare L.) Boron tolerance mechanism. Dissertation, Fatih University

Tombuloglu G, Tombuloglu H, Sakcali MS, Unver T (2015) High-throughput transcriptome analysis of barley (Hordeum vulgare) exposed to excessive boron. Gene 557:71-81

Trapnell C, Pachter L, Salzberg SL (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515

Troncoso-Ponce MA, Kilaru A, Cao X, Durrett TP, Fan JL, Jensen JK, Thrower NA, Pauly M, Wilkerson C, Ohlrogge JB (2011) Comparative deep transcriptional profiling of four developing oil seeds. Plant J 68:1014–1027

Vega-Arreguin JC, Ibarra-Laclette E, Jimenez-Moraila B, Martinez O, Vielle-Calzada JP, Herrera-Estrella L, Herrera-Estrella A (2009) Deep sampling of the Palomero maize transcriptome by a high throughput strategy of pyrosequencing. BMC Genomics 10:299

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487

Villar E, Klopp C, Noirot C, Novaes E, Kirst M, Plomion C, Gion JM (2011) RNA-seq reveals genotype-specific molecular responses to water deficit in eucalyptus. BMC Genomics 12:538. doi:10.1186/1471-2164-12-538

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, Landherr L et al (2009) Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10:347

Wang W, Vinocur B, Altman A (2003) Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. Planta 218:1–14

Wang W, Wang Y, Zhang Q, Qi Y, Guo AD (2009) Global characterization of *Artemisia annua* glandular trichome transcriptome using 454 pyrosequencing. BMC Genomics 10:465

Wang WX, Vinocur B, Shoseyov O, Altman A (2001) Biotechnology of plant osmotic stress tolerance: physiological and molecular considerations. Acta Horticult 560:285–292

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweet potato (*Ipomoea batatas*). BMC Genomics 11:726

Weber AP, Weber KL, Carr K, Wilkerson C, Ohlrogge JB (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. Plant Physiol 144:32–42

Xu P, Liu Z, Fan X, Gao J, Zhang X, Zhang X, Shen X (2013) De novo transcriptome sequencing and comparative analysis of differentially expressed genes in *Gossypium aridum* under salt stress. Gene 525(1):26–34

Ye J, Fang L et al (2006) WEGO: a web tool for plotting GO annotations. Nucleic Acids Res 34(Web Server issue):W293–W297

Zagrobelny M, Scheibye-Alsing K, Jensen NB, Moller BL, Gorodkin J, Bak S (2009) 454 Pyrosequencing based transcriptome analysis of *Zygaena filipendulae* with focus on genes involved in biosynthesis of cyanogenic glucosides. BMC Genomics 10:574

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821–829

Zhang JJ, Zhoud ZS, Songe JB, Liud ZP, Yanga H (2012) Molecular dissection of atrazine-responsive transcriptome and gene networks in rice by high-throughput sequencing. J Hazard Mater 219–220:57–68

Zhu JK (2001a) Plant salt tolerance. Trends Plant Sci 6:66–71

Zhu JK (2001b) Cell signaling under salt, water and cold stresses. Curr Opin Plant Biol 4:401–406

Zhu JK (2002) Salt and drought stress signal transduction in plants. Annu Rev Plant Biol 53:247–273

Chapter 13
CNV and Structural Variation in Plants: Prospects of NGS Approaches

Enrico Francia, Nicola Pecchioni, Alberto Policriti, and Simone Scalabrin

The present chapter focuses on copy number variants (CNVs). It firstly summarizes how CNVs are classified within structural variants (SVs), which are the mechanisms causing their onset, and to which extent they have been discovered in plant genomes. Moreover, as most of the CNVs reported so far overlap with protein-coding sequences and result in gains and losses of gene copies that might have a straight influence on gene/transcript dosage (Chia et al. 2012), particular attention is given to the role played by copy number variation (CNV) in the regulation of relevant adaptive traits, e.g. plant development, as well as resistance to abiotic stresses. A full range of structural variation could thus be detected from next-generation sequencing (NGS) data, including translocations, and CNVs (for a review, see Abel and Duncavage 2013). However, the complexity of plant genomes and the short read length obtained from NGS platforms pose new bioinformatic challenges associated with their detection. After the discussion about the computational issues, the array of available methods for CNV discovery from NGS data is reviewed. Notably, although numerous

E. Francia, Ph.D. • N. Pecchioni, Ph.D. (✉)
Department of Life Sciences, University of Modena and Reggio Emilia, via Amendola, 2, Reggio Emilia 42122, Italy

CGR – Center for Genome Research, University of Modena and Reggio Emilia, Via Campi, 287, Modena 41126, Italy
e-mail: nicola.pecchioni@unimore.it

A. Policriti, Ph.D.
Department of Mathematics and Computer Science, University of Udine, Udine 33100, Italy

IGA - Institute of Applied Genomics, Parco Scientifico e Tecnologico "L. Danieli", Udine 33100, Italy

S. Scalabrin, Ph.D.
Department of Mathematics and Computer Science, University of Udine, Udine 33100 , Italy

IGA Technology Services, Parco Scientifico e Tecnologico "L. Danieli", Udine 33100, Italy

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_13

software packages are available for NGS analysis, there is currently no single informatic method capable of identifying the full range of structural DNA variation, and multiple complementary tools are required for robust CNVs detection. Finally, future bioinformatic and applicative prospects for such genomic variants are discussed.

Copy Number Variation Is Part of Genome Structural Variation

Plant nuclear genomes display extensive variation in size, chromosome and gene number, and number of genome copies per nucleus (Kellogg and Bennetzen 2004). Such genomic variability can be present in many forms, including single nucleotide polymorphisms (SNPs), variable number of tandem repeats (VNTRs; e.g., mini- and microsatellites), presence/absence of transposable elements (e.g., retrotransposons and DNA transposons), and different forms of structural variation (SV) (Fig. 13.1). On the basis of their nature, SVs are classified in (1) chromosomal inversions when a segment of a chromosome is reversed end to end, (2) translocations in which rearrangements of parts of non-homologous chromosomes are involved, and (3) CNVs. Scherer (2007) masterly overviewed how descriptors of variation began in the realm of cytogenetics in the 1960s and in the 1970s, continued in the field of molecular genetics and, most recently, in that of cytogenomics, which bridges the gap for detection of genomic variants. Owing to Feuk et al. (2006), and as said in the introductory paragraph, SV should cover by definition the genomic variation that affect large DNA segments, ranging from 1 kb to several Mb ("submicroscopic" size). The designation of the category "1 kb to submicroscopic" is somewhat arbitrary at both ends, but is used for operational definition. In a broad sense, structural variation has been used to refer to genomic segments both smaller and larger than the narrower operational definition. CNVs are currently defined as unbalanced changes in the genome structure and represent a large category of genomic structural variation, which according to Alkan et al. (2011) should include by definition insertions (i.e., the addition of one or more base pairs into a DNA sequence), deletions (i.e., the loss of any number of nucleotides, from a single base to an entire piece of chromosome), tandem or interspersed duplications (i.e., any duplication of a region of DNA). According to these authors, also the single base INDELs should be ideally ascribed to CNVs. NGS, in conjunction with increasingly powerful bioinformatic tools, made possible the identification of polymorphic regions of >50 bp in size, traditionally defined as INDELs, that could be included among SVs (Alkan et al. 2011). In other reports (e.g., Springer et al. 2009), the definition of CNV is associated with that of presence-absence variation (PAV), that should include the insertions and deletions distinct from the typical CNVs; to which, in a restrictive view, should only be ascribed duplications. In the present chapter we prefer to embrace the wide-angle vision of CNVs, by including present-absent variants (PAVs) into this group of SVs. Then, in accordance with most literature reports, we also prefer to exclude from CNVs the structural variations <1 kb (Fig. 13.1). CNV sometimes exhibits strong associations with specific biological functions.

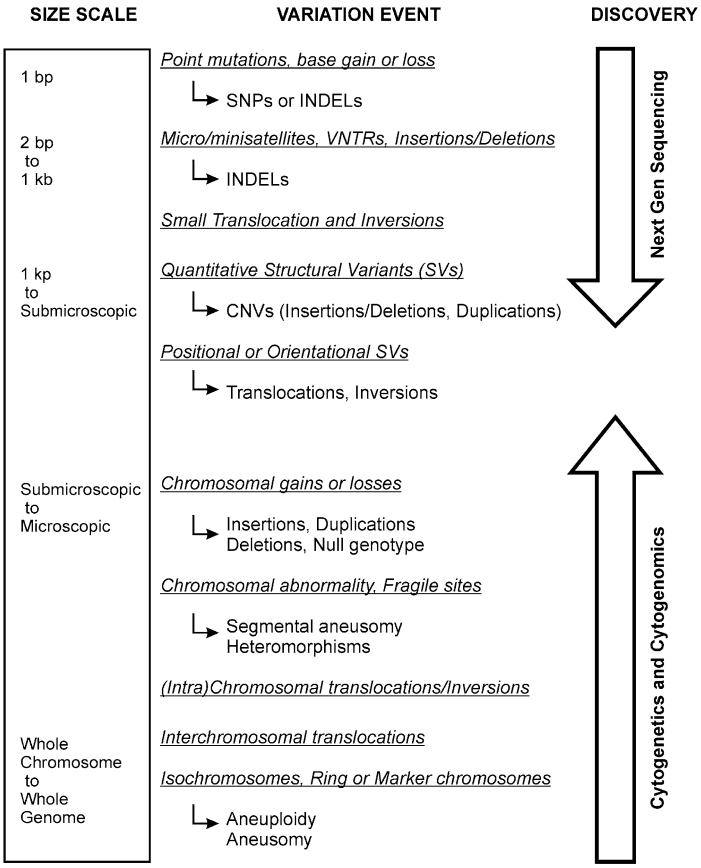

Fig. 13.1 General diagram of genomic structural variation (*SV*). Events ranging from single-base sequence variation to either whole chromosome or genome changes are *italics underlined* and ordered in the figure according to their physical size (*rectangle* to the left). Major categories of single nucleotide variants (*SNVs*) and structural variants (*SVs*) resulting from variation events are indicated by *solid arrows*; SVs <1 kb were excluded from CNVs in accordance with most literature records (see text for details). *Open arrows* show the general approaches applied for discovery the different variation events. Figure modified from Scherer et al. (2007)

Structural variation is therefore increasingly recognized, in humans as well as in other organisms, as a common feature and evolutionary force of genomes, where CNVs and associated gene dosage effects have been implicated in a number of trait/phenotypes (Girirajan et al. 2011; Cantsilieris and White 2013).

Diffusion of CNVs Within Genomes

Owing to the biomedical focus of most studies, at present the best data on CNVs come from human genome. Great interest in CNVs was stimulated by two initial papers of Iafrate et al. (2004) and Sebat et al. (2004) in the early 2000s. Both these

papers described large-scale copy number polymorphism in the human genome. CNV was found surprisingly common among humans. For example, an early study by McCarroll et al. (2008) has revealed that a sizable proportion—from 175 to 230 autosomal loci spanning approximately six megabases—of the human genome varies in copy number between two unrelated individuals. Considering only protein-coding genes, studies show that any two humans are likely to differ at CNVs completely encompassing approximately 105 genes. Interestingly, a considerable higher gene CNV was found in maize (Swanson-Wagner et al. 2010); however, the importance of this result cannot be overemphasized: any two individual genomes taken from nature, in any species, will have dozens to hundreds of differences in their total number of functional genes. Currently, it is estimated that common CNVs occur in approximately 10–12 % of the human reference genome (Conrad et al. 2010; Redon et al. 2006). In human genome CNVs are often detected in regions that contain protein-coding genes or important regulatory elements. CNVs may also affect gene regulation by position effects, and CNVs that partially overlap a gene sequence may disrupt the structure of the gene and impair its function (for a review see Zmieńko et al. 2014). In a comparison study between humans and chimpanzee, beside a conservation of many CNV regions between the two species, some of these regions appeared to be "hotspots" for the genesis of this kind of variation (Perry et al. 2006). CNVs in plants have not been so thoroughly studied, notwithstanding the significant number of diverse fully sequenced genomes since 2000. It is only in the last 5 years that CNVs have attracted the attention of plant biologists and geneticists, likely stimulated by the first findings of association to phenotypes in 2009 and 2010, and leading to the first estimates of the extent of CNV in plant genomes. Notably, in plant genetics, the individual organisms are mainly treated as representatives of one of the following sub-types: (a) cultivars (also named varieties), which are distinct, often intentionally bred subsets of a species that will behave uniformly and predictably when grown in the environment to which they are adapted or (b) accessions, which are collections of plant material from a particular location that are given unique identifiers. Accordingly, CNVs in plants are often recognized and discussed as polymorphisms distinguishing cultivars/accessions of one species rather than affecting individual plants (Chia et al. 2012; Cao et al. 2011; Xu et al. 2012). The crop plant in which CNVs have been primarily investigated, and for which exist the deepest knowledge is maize—*Zea mays* L. (Springer et al. 2009; Swanson-Wagner et al. 2010; Beló et al. 2010; Jiao et al. 2012). After the release of the complete genome sequence of the inbred line B73, the extremely high genomic diversity exhibited by maize has become accessible at a level of detail never had before. Several studies revealed extensive structural variation, including hundreds of CNVs and thousands of the cited PAVs. Basing on comparative genomic hybridization (CGH) Springer et al. (2009) and Beló et al. (2010) detected thousands of dispersed as well as clustered CNVs in the maize genome, between B73 and Mo17 inbred lines or among 13 inbreds compared to B73, respectively. Two main factors affected the estimation of the number of CNVs detected between different inbreds. First, the microarray platform used was primarily developed for gene expression with not uniform distribution of genes along the maize genome (e.g., with fewer probes in

the paracentromeric regions). Second, the majority of the probes were designed to be complementary to the B73 allele, and therefore sequences absent from B73 could not be detected. As a consequence, the number of CNVs identified was underestimated, especially with respect to small CNVs, as the methodology favors detection of large insertion–deletion variants. However, the high level of structural variation and differences in genome content observed in maize are unprecedented among higher eukaryotes. Lai et al. (2010) characterized genetic variation in the six elite strains most commonly used to make commercial hybrids. As already hypothesized by Springer et al. (2009), the authors discussed the potential roles of complementation of gene PAVs, CNVs, and other mutations in contributing to heterosis. Swanson-Wagner et al. (2010) analyzed structural variation between diverse maize inbreds and inbred wild teosinte lines, providing evidence for widespread genome content variation. Over 70 % of the CNV/PAV examples were identified in multiple genotypes, and the majority of events were observed in both maize and teosinte, suggesting that these variants predate domestication and that it seems not having been strong selection acting against them. Partially in contrast with this observation, Jiao et al. (2012) reported extensive CNVs occurring through the maize breeding history. By sequencing of 278 inbred lines from different periods of breeding history, including deep resequencing of 4 lines with known pedigree information, these authors could conclude that, even within identity-by-descent regions, extensive variation caused by SNPs, INDELs, together with CNVs occurred quite rapidly during breeding. In particular, 8.5 % of maize genes showed CNV among the four compared genomes, and an average CNV rate was calculated, although lower for maize compared to that described in humans (8.57×10^{-4} per gene per year vs. 1.2×10^{-2}) (Jiao et al. 2012). As a second important crop surveyed for CNVs, soybean reference genome of cultivar Williams 82 has been compared with introgressed regions from parent Kingwa by analyzing nucleotide and structural differences between Williams 82 individuals (Haun et al. 2011). The authors found that in soybean the impact of intracultivar genetic heterogeneity can be significant, with a high rate of structural and gene content variation and, as hypothesized in humans, the presence of conspicuous CNV hotspots. McHale et al. (2012) combined and compared two approaches for the evaluation of genome-wide structural and gene content variation among four soybean genotypes: microarray CGH and exome DNA capture and resequencing. As an interesting result of the analyses, the regions most enriched for SVs were gene-rich regions harboring clusters of multigene families. Only members of multigene families that are located within clusters tend to be associated with CNV regions. Among these multigene families, the most abundant were the nucleotide-binding and receptor-like classes, presumably important for plant defense against pathogens. In terms of CNV distribution, soybean showed relatively long chromosomal regions (and nearly entire chromosomes) that exhibit virtually no SV among genotypes, interspersed with pockets of high SV ranging from several kb to greater than 10 Mb in length. By resequencing and comparing two sweet and one grain sorghum (*Sorghum bicolor* L.) inbred lines to the reference accession BTx623, Zheng et al. (2011) came to similar result. Along with INDELs PAVs and SNPs, more than 17,000 CNVs (>2 kb in length) were retrieved. While the majority of the

large-effect structural variations resided in genes containing LRR, PPR repeats and in disease resistance R genes, annotation analysis showed that 2,600 genes had 3,234 CNVs, and 32 genes had CNVs in all three sorghum lines (Zheng et al. 2011). The first catalog of CNVs in a diploid Triticeae species has been reported by Muñoz-Amatriaín et al. (2013) for the barley (*Hordeum vulgare* L.) crop. The authors developed a CGH array covering approximately 50 Mb of repeat-masked sequence of the reference cv. Morex and compared via genomic hybridization a collection of 14 genotypes including eight cultivars and six wild barleys. Almost 15 % of all the sequences considered were affected by CNV and more than 60 % were found in two or more genotypes. As already observed in the maize genome (Springer et al. 2009; Swanson-Wagner et al. 2010; Beló et al. 2010) CNVs in barley are enriched near to chromosome ends, apart in one chromosome (4H), that showed the lowest frequency of CNVs. CNV affects 9.5 % of the coding sequences represented on the array and, similarly to what observed in soybean, the genes affected by CNV are enriched for sequences annotated as disease resistance proteins and protein kinases. The list of agriculturally relevant species surveyed for presence of CNVs extends at least to allotetraploid wheat (Saintenac et al. 2011), rice (Yu et al. 2013), and tomato (Causse et al. 2013). A significant presence of such SVs has been verified consistently in all the three species. By a sequence capture assay restricted to 3.5 Mb exon regions, for a total of 3,497 genes of tetraploid wheat compared between cultivar Langdon and a wild emmer accession, Saintenac et al. (2011) found 85 CNV targets; among these, 77 variants were due to an elevated number of copies in the Langdon genome and only 8 variants resulted from copy increase in the wild emmer genome. In the rice CGH study, Yu et al. (2013) identified 2.69 % of rice genome interested by CN variable regions (CNVRs), overlapping 1,321 genes, these significantly enriched for cell death, protein phosphorylation, and defense response, as already observed in soybean and barley. The 1,686 putative CNV regions identified in tomato impacted a total of 1,235 genes, with significant differences between the eight resequenced genotypes, and cell death process genes represented in significant excess (Causse et al. 2013).

Mechanisms Leading to Variation in Number of Copies

As a general rule, alteration in copy number involves change in the structure of the chromosomes such that two formerly separated DNA sequences are joined together. Several mechanisms have been postulated to explain the formation and then the variation in number of copies of CNVs (Hastings et al. 2009a). However, the mechanisms of all structural changes that involve chromosomal DNA are substantially the same, and occur by two general mechanisms: homologous recombination (HR) and non-homologous recombination (NHR). HR is a complex process whereby DNA segments that share significant sequence homology are exchanged. This definition entails the requirement for broad DNA sequence identity; however, in yeast it is thought that as little as 30 bp are sufficient (Haber 2000). In plants, a few hundred

base pairs can engage the HR machinery (Puchta and Hohn 1991), but it is still unclear whether there is a lower limit, nor what is the dependence on the type of partners (Lieberman-Lazarovich and Levy 2011). Sequence microhomology (i.e., very few bases of identity) or no homology are instead the basic events for NHR. Although HR provides vital repair mechanisms, meiosis requires crossing over and a possible side effect of this requirement is the rather high frequency of CNVs produced—according to the estimates reported by Lupski (2007). According to this author, such frequency ranges from 10^{-6} to 10^{-4} copy number changes per gamete. Several mechanisms are based on HR for repairing DNA breaks and gaps; among these, the best studied is called double-stranded break (DSB)-induced recombination. Owing to previous research done in *Saccharomyces cerevisiae* one of DSB repair models (namely, synthesis-dependent strand annealing—SDSA), which does not generate crossovers, could produce variations in copy number when the DNA template contains direct repeats (for a review see, Pâques and Haber 1999). A more important HR mechanism is the non-allelic homologous recombination (NAHR), between DNA segments on the same chromosome and of high similarity, but that are not alleles. NAHR usually involves low-copy repeats (LCRs)—DNA segments larger than 1 kb that are generated during ancient duplication events. Depending on the LCR location, NAHR can lead to intrachromatid, interchromatid, or interchromosomal rearrangements. The type of rearrangement depends on LCR orientation: the repeats may be direct, opposite or mixed. The orientation determines whether NAHR leads to the deletion, reciprocal duplication, or inversion of the DNA segment flanked by the LCRs. In maize, some transposon elements have been shown capable of directly inducing tandem sequence duplications, and let to hypothesize that this activity has contributed to the evolution of the maize genome (Zhang et al. 2013). Besides repairing two-ended DSBs, HR can repair collapsed or broken replication forks in a process called break-induced replication (BIR). Several authors discussed the possible involvement of BIR in a microhomology-mediated mechanism of copy number change (Hastings et al. 2009b). Finally, a minor HR player in the formation of CNVs is a DSB mechanism known as single-strand annealing (SSA). In yeast, SSA has been found responsible for deletions of up to a few tens of kb (Pâques and Haber 1999), while in plants SSA can lead to efficient sequence deletions between direct repeats and this might, for example, explain the accumulation of single long terminal repeats of retroelements in cereal genomes (Puchta 2005).

Concerning NHR, other mechanisms of DSB repair either do not require homology or need very short micro-homologies for DNA repair: non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), and breakage–fusion–bridge cycle (Puchta 2005). All these phenomena increase the probability of genetic changes such as CNV. Another potential non-homologous mechanism is fork stalling and template switching (FoSTeS). FoSTeS is caused by DNA replication errors when replication forks stall, in a manner in which the 3′ primer end of a DNA strand can change templates to an ssDNA template in a nearby replication fork (Lee et al. 2007). FoSTeS events may generate insertions, deletions, or more complex rearrangements such as CNVs (Lee et al. 2007). However, Hastings et al.

(2009b) proposed a new model—microhomology-mediated break-induced replication (MMBIR)—that in addition to the events included in FoSTeS could also lead to translocations. Interestingly, MMBIR supports the hypothesis of an increase in the frequency of CNVs produced when cells are under stress. This observation from molecular evidences is consistent with the intriguing hypothesis of an adaptive evolutionary value of CNV, when organisms are challenged by environmental stresses (see below). Such switch from high-fidelity to error-prone DSB repair in stress conditions seems common from bacteria to plants (DeBolt 2010). Finally, it must be underlined that CNVs are not randomly distributed in genomes, but tend to be clustered in CNV hotspots (Perry et al. 2006), in regions of complex genomic architecture. There is therefore ample evidence that specific features of chromosomal architecture are also involved in CNV generation, and this entails that multiple genomic features can affect the probability of CNV occurrence.

Do CNVs Have a Biological Meaning?

Association to Phenotypes

Since the 1980s it is known that the human genome contains apart from single base and short repeat polymorphisms another abundant source of variation, involving deletions, insertions, duplications, and complex rearrangements. Nevertheless, the first evidence of a phenotypic role of CNVs has come with the elucidation of the etiology of Charcot–Marie–Tooth neuropathy type 1A, due to gene duplication rather than to point mutations (Lupski et al. 1991). Since 1991 and until 2006, with the scientific world fully dedicated to the exploitation of SNP-associated traits, only a small number of pioneer studies advanced knowledge of CNV impacts on human diseases, before the systematic characterization of Redon et al. (2006). They identified CNVs covering approximately 12 % of the human genome, and hypothesized potential alterations of gene dosage, gene disruption or perturbed regulation of their expression, even at long-range distances. After the first global searches aimed to discover and catalog these structural variations in the human and mouse genome, an array of different experiments mostly performed as case–control studies allowed to characterize an increasing number of CNV-associated phenotypes (diseases) in humans, such as Crohn's disease (McCarroll et al. 2008). Diskin et al. (2009) demonstrated, in a disease for which SNP variations are known to influence susceptibility, that CNV at 1q21.1 is associated with neuroblastoma and implicates a novel gene in early tumorigenesis. Sometimes, genetic risk factors have been missed because association studies have sought risk-associated SNPs, while ignoring structural variation causing gene copy number changes. This is the case of CNVs associated with colorectal adenoma recurrence (Laukaitis et al. 2010). But it is in particular with many developmental neuropsychiatric disorders that rare CNVs have unprecedented levels of statistical association. These CNV-associated disorders include schizophrenia, autism spectrum disorders, intellectual disability, and attention

deficit hyperactivity disorder (ADHD); however, as CNVs often include multiple genes, causal genes responsible for CNV-associated diagnoses and traits are still poorly understood (Hiroi et al. 2013). Among these associations, the 16p11.2 copy variant phenotype of neurocognitive defects was found to be driven by the KCTD13 gene dosage changes within the CNV region encompassing 29 genes (Golzio et al. 2012). As regards other investigated traits, finally and interestingly, severe obesity and being underweight could be mirror extreme phenotypes of the same CNV at 16p11.2 locus, respectively, associated with a large (600 kb) deletion vs. a duplication of the region (Jacquemont et al. 2011).

The intuitive scientific question whether CNVs can modify gene expression is a key issue for their association to phenotypes where differential gene expression plays a role. The majority of experiments found out that not only variations in gene copy numbers can modify gene expression in carrier genotypes, but importantly they can also significantly influence expression time courses. In a global survey in humans, Stranger et al. (2007) observed that CNVs captured a significant percentage of the total genetic variation in gene expression, 17.7 %, although lower than the remaining part attributed to SNPs (83.6 %). In a study throughout mouse development, Chaignat et al. (2011) observed that CNV genes are significantly enriched within transcripts showing variable time courses between mice strains; thus, modifications of the copy number of a gene may alter not only gene expression, but also potentially alter timing of its expression. Henrichsen et al. (2009) found that not only expression of human genes within CNVs tend to correlate with copy number changes, but also that CNVs can influence the expression of close genes, with an effect extending in the vicinity up to a distance of 0.5 Mb; moreover, they can also have a global influence on transcriptome. An intriguing effect on gene expression has been shown by a promoter competition between copy number variant α-globin genes and the NME4 gene, located 300 kb apart from the α-globin cluster, for which the deletion of two α-globin genes is unlocking higher NME4 expression by a regulator (Lower et al. 2009).

Also in plants, SV has been hypothesized to be a driving force behind phenotypic variation (Chia et al. 2012). First clear associations to phenotypes in plants follow at distance the discoveries made in human genetics, with the first report in barley dating 2007. The boron-toxicity tolerant cultivar Sahara contains about four times as many *Bot1* boron transporter gene copies compared to intolerant genotypes, and produces significantly more *Bot1* transcripts. *Bot1* transcript levels identified in barley tissues are consistent with an avoidance strategy, by limiting the net entry of boron into the root and by disposing boron from leaves via hydathode guttation (Sutton et al. 2007). A very similar genetic strategy has been observed recently in maize for tolerance to Aluminum, i.e. to acidic soils. The expansion in MATE1 (multidrug and toxic compound extrusion 1) copy number is associated with higher MATE1 expression, which in turn results in superior Al tolerance; the three MATE1 copies in the rare tolerant genotypes (all containing three copies) are identical and are part of a tandem triplication, absent in the vast majority of susceptible accessions that carry a single copy of the transporter gene (Maron et al. 2013). The frost-tolerant barley cultivar Nure contains tandem segmental duplications through the

CBF2A-CBF4B genomic region of the CBF gene cluster on chromosome 5H, that differentiate freeze-tolerant from sensitive genotypes, which carry single copies of those genes. The higher copy number of CBF genes is associated with higher gene expression in tolerant genotype Nure of the transcription factors under short days (Knox et al. 2010). Although observed for an effector gene, at the end of the final response cascade to cold, CNV of Y_2K_4 dehydrin in *Medicago* has been hypothesized as a duplication of dehydrin genes in cold-tolerant cultivated alfalfa genotypes (Castonguay et al. 2013). In a review of Oh et al. (2012) about tolerance of plants to extreme conditions, gene duplication is indicated as one of three possible strategies to cope with extreme abiotic stress conditions. Among the examples reported to support the hypothesis, HKT1, a plasma membrane Na^+/K^+ transporter considered to be a genetic determinant of salt tolerance, exists as tandem duplicated copies in two salt-tolerant *Thellungiella* species. As a second example, the duplication of NHX8 homologs, known to encode a putative Li^+ transporter in *A. thaliana*, leads to a constitutively higher expression in *Thellungiella parvula* than in *A. thaliana*, and this in turn might be responsible for the apparently enhanced tolerance of *T. parvula* to high Li^+ in its natural habitat. In maize, a recent genome-wide SNP screen of 103 diverse maize and teosinte lines (Chia et al. 2012) suggests a correlation between genomic regions containing structural variation – detected as read-depth variants (RDVs) in genome resequencing – and QTLs for agronomic traits. As an interesting example, genomic regions containing QTLs for leaf architecture and resistance to northern and southern leaf blight are enriched for RDVs. This suggests a potential role for CNV/PAV in generating phenotypic variation for these agronomic traits. Schnable and Springer (2013) hypothesize a generic role for gene CNV to help explaining heterosis. In fact, complementation of allelic variation, as well as complementation of variation in gene content and expression patterns, is likely to be important contributors to this trait of paramount importance in maize. CNV/PAV has been reported to be differentially represented among genes categorized as being involved in stress and stimulus response, perhaps in part because this category includes some large gene families (e.g., NBS-LRR genes). This pattern is detectable on a genome-wide scale in maize (Chia et al. 2012), rice (Xu et al. 2012) and in other plants. An interesting example of multiple resistance genes acting by means of a structural variation is *Rhg1* nematode resistance QTL in soybean. Cook et al. (2012) demonstrated how this resistance is governed by a peculiar CNV of multiple genes. Ten tandem copies of the 31-kilobase segment identifying the *Rhg1* locus are present in an rhg1-b resistant haplotype vs. one copy per haploid genome in susceptible varieties. In this multigene segment, overexpression of the individual genes was ineffective, but overexpression of the genes together conferred enhanced soil cyst nematode resistance. Hence, SCN resistance mediated by the soybean quantitative trait locus *Rhg1* is conferred by CNV that increases the expression of a set of dissimilar genes in a repeated multigene segment.

Regulation of plant development is the last group of plant phenotypic traits that are being increasingly associated with CN variations. In barley, Nitcher et al. (2013) demonstrate that the *HvFT1* (FLOWERING LOCUS T homolog, corresponding to the VRN-H3 locus) allele present in the barley accession BGS213 and associated

with a dominant spring growth habit, carries at least four identical copies of *HvFT1*, whereas most barley varieties harbor a single copy. The increased copy number is associated with earlier transcriptional up-regulation of *HvFT1*, thus giving further support to the hypothesis made in humans that CNV is not only leading to differences in gene expression, but also to differences in expression time course. In wheat, two key regulators of flowering in response to light and temperature have been found to be ruled by CNV associated with altered gene expression. Alleles with an increased copy number of photoperiod response gene *Ppd-B1* confer an early flowering day neutral phenotype and have arisen independently at least twice. At the same time, plants with an increased copy number of vernalization requirement gene *Vrn-A1* have an increased requirement for vernalization so that longer periods of cold are required to potentiate flowering (Díaz et al. 2012). The results shed new light on regulation of flowering in wheat, and intriguingly suggest that CNV plays a significant role in wheat and plant adaptation.

Evolutionary and Adaptive Value of CNVs

As stated by Schrider and Hahn (2010) and by Kondrashov (2012), although it might be too early to tell whether or not a substantial fraction of gene copies have initially achieved fixation in eukaryotes by positive selection for increased dosage, nevertheless enough examples have accumulated in the literature to strongly suggest an adaptive value for such genetic variation. As a consequence of this, a complete understanding of the molecular basis for adaptive natural selection must necessarily include the study of copy number variation. One of the clearest examples supporting such hypothesis comes from budding yeast (Stambuk et al. 2009). In five industrially important *S. cerevisiae* strains responsible for the production of fuel ethanol from sugarcane, there have been found significant amplifications of the telomeric SNO and SNZ genes, which are involved in the biosynthesis of vitamins B6 (pyridoxine) and B1 (thiamin), and confer the ability to grow more efficiently under the repressing effects of thiamin, especially with high sugar concentrations. These genetic changes have likely been adaptive and selected for in that specific industrial environment. Similar effects of the feeding environment on CNV were observed in wood decaying fungi, where CNV was observed in members of the detoxification pathways belonging to multigenic families such as the cytochrome P450 monooxygenases and the glutathione transferases, as an adaptive strategy allowing these basidiomycetes to deal with the plethora of potential toxic compounds resulting at least partly from wood degradation (Morel et al. 2013). In humans, the AMY1 α-amylase gene, which encodes a protein catalyzing starch degradation constitutes an interesting example. It has been found a gene copy number three times higher in humans compared to chimpanzees, and higher expression levels of salivary amylase protein, suggesting that humans were favored in the gene dosage due to an increase of starch consumption in their evolutionary history (Perry et al. 2007). As pointed out by Bailey et al. (2008), in a global survey of human copy

number genes, many examples of gene CNVs described within the human population due to their association with phenotype and disease, also before the NGS era, can be postulated to have played important roles in human adaptation to changing environmental conditions and infectious pathogens.

In plants, the common observed association between abiotic and biotic stress tolerant phenotypes and gene CNV is coupled to the observation in Arabidopsis (DeBolt 2010), although common to all organisms (Hastings et al. 2009a; Freeman et al. 2006), that CNVs form at a faster rate than other types of mutation. A striking example of such a faster rate is the generation of significant numbers of CNVs in Arabidopsis lineages after only five generations under low and high temperature and chemical (salycilic acid spray) stresses, with positive selection for fecundity, while genotypes deriving from the same mother plants by selfing did not display any differences in CNV when growing under normal conditions (DeBolt 2010). Boyko and Kovalchuk (2011), from their previous experiments about signaling in plant–pathogen interactions, hypothesize the generation in plants infected with a compatible pathogen of a systemic recombination signal (SRS) that precedes the spread of pathogens and results in an increase of the somatic and meiotic recombination frequency. Although yet to be fully validated, the hypothesis is an intriguing further support to a wide environmental adaptive role for the origin of SVs. In a very interesting review about genetic variation in extremophile plants (adapted to extreme environmental conditions), Oh et al. (2012) argue that there is little overall evidence that polyploidy itself is a major evolutionary driving force leading to extremophiles, while tandem duplications seem to have a more important role in shaping genomes for stress adaptations. The evolutionary meaning of local gene duplications could be in fact viewed also in comparison with polyploidy, common in plants, and for a long time considered as a main evolutionary driver in these organisms. In humans, Makino and McLysaght (2010) observe that duplicated genes deriving from two ancestral WGD (whole genome duplication; i.e. ohnologs) have rarely experienced subsequent small-scale duplication (SSD), are refractory to CNV, are dosage-balanced and preferably retained in human populations; by contrast, genes that have experienced SSD are more likely to also display CNV and dosage unbalance. Similar observations in plants took Birchler (2012) to conclude that different fates can be observed for duplicate genes depending on whole genome or segmental duplication. Following polyploidy formation, members of macromolecular complexes persist in the evolutionary lineage longer than random genes, while a complementary pattern is found for segmental duplications in that there is an underrepresentation of members of macromolecular complexes.

What written about adaptive value of CNVs is mostly referred to examples of copy number variable genes, and the majority of validated phenotypes present in the literature refer to these cases. However, the case of the relatively large structural variation at *Rhg1* locus in soybean suggests that also other more complex copy unbalances in higher organisms, if at similar faster mutation rates, can be included within the same evolutionary meaning.

NGS Approaches and Bioinformatic Tools for CNV Detection

In this section, we are going to discuss some bioinformatics issues involved in the discovery and classification of SV, with special emphasis on CNV in plants. We consider these issues trying to answer the following questions:

1. general: Given the mathematical definitions of the problems we want to solve, what are the main computational bottlenecks to face and what kind of limits can we put to the (abstractly obtainable) answers?
2. practical: On the grounds of the given definitions, which ones among the concrete solutions proposed in the literature—and to what extent—reach the potential frontiers of implementable tools?
3. technological: Is the interplay among proposed definitions, computational problems, and available (or foreseeable) technologies for data production, going to change significantly the landscape in the (near) future?

We will see that the search of variations among genomes of different organisms of the same species is a challenging subject, as a result of the difficulties involved in answering each one of the above questions. The problem is mathematically elusive, as a precise definition is either quickly unrealistic or impossible to satisfy; practical solutions proposed are often difficult to judge or classify, because of the large amount of specific and rapidly changing sets of heuristics implemented. It is often not clear how the technological changes that we expect will take place in the near future, will modify the amounts and the kind of data that soon will be available for analysis. Nevertheless, the bioinformatics aspects involved in the field make the challenges exciting, as it is clear that only a coordinated effort towards a clear specification and a compilation of realistic needs can result in the design of a new generation of useful tools.

The Computational Problem

From a computational point of view, we begin by attempting a classification of SV and CNV. Any classification must assume the existence of a reference genome G for the organism under study. The reference can be either the first (or most reliable) available sequenced genome for the species, or a core genome resulting as a common factor of previous analyses. The first class of objects (SV) is usually defined as the collection of sub-sequences σ that may or may not appear in G. In such terms, SVs include *any* possible variation to search and classify. Among SVs we can isolate CNVs as those sub-sequences γ whose characteristic feature is presence/absence together with the number of their occurrences. As we pointed out, any sensible definition is to be given with respect to a genome sequence to be considered our fixed reference system. This is true even in cases in which an "official" reference is not available and a comparative study between two or more individuals is carried out: in these cases, the reference is fixed on-the-fly but is, however, present. Hence, for example, presence in the individual under study and absence in the reference corresponds to a number of occurrences equal to one against a number of

occurrences equal to zero (infinite ratio). In general, when a γ occurs in the reference we can talk about the rate of its occurrence. For both SVs and CNVs, the definition should be further refined by (at least) specifying:

- the (lower) limits in length for σ's and γ's, thereby introducing a finer classification on both categories;
- the number and kind of allowed alignment errors, while establishing presence/absence or evaluating the number of occurrences.

The above classification cannot be rigid: two shorter sub-sequences cannot be considered equal by the same percentage of errors (mismatch, insertion/deletions of characters) employed for significantly longer ones. Moreover, even though CNVs do change the total length of the genome, a detection based on a variation of the total length is not of any practical use. On the ground of the grid defined above, we can then finally enter within a more functional analysis of the sub-sequence considered. Each σ or γ can be classified on special patterns defining its encoding, compositional, or otherwise syntactically characterizing feature.

NGS and the Main Techniques of CNV Discovery

Historically, two general categories of methods were used to detect CNVs and regions with overlapping CNVs (CNVRs): array-based comparative genome hybridization (CGH) and reference genome-based NGS. The first ("hybridization-based mapping") followed the observation that any region duplicated or deleted in an individual sample will show an excess or deficit, respectively, of DNA that is highly similar to that region relative to the reference genome. These methods were therefore aimed at detecting these localized differences in relative DNA content. The second category of methods (sequencing) does not detect the duplications and deletions directly, but instead detects length differences in the size of captured fragments from a sample relative to the reference genome. Fragments that appear too large must contain insertions or duplications, while those that are too small must contain deletions. Other methods, such as quantitative PCR (D'haene et al. 2010) and fluorescent in situ hybridization (Cook et al. 2012), can be used to verify CNVs but they are generally not useful for the discovery process. The current approach for CNV discovery uses NGS high-throughput DNA sequencing technology. This approach has been proven effective for the discovery and mapping of SVs at nucleotide resolution in plants, animals and humans (Cao et al. 2011; Daines et al. 2009; Yoon et al. 2009; Mills et al. 2011; Bickhart et al. 2012).

A Classification of NGS Technologies

The previously used array-based methods could still provide a cost-effective mean for CNV discovery but they suffer of low throughput and low resolution of breakpoints, in the best cases hundreds of bp (Conrad et al. 2010; Park et al. 2010).

Precise characterization of breakpoints, which may capture the signature of potential mutational mechanisms, is crucial for designing robust genotyping assays and assessing the functional content of detected CNVs (Li and Olivier 2013). Moreover, these methods are limited to sequence present in the reference assembly used to design the probes and they cannot neither identify balanced structural variations nor specify the location of a duplication (Alkan et al. 2011). In order to overcome the above problems sequencing has been used in the last years. Initially only Sanger sequencing (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008; Korbel et al. 2009) was used, then also Second (Bentley et al. 2008; Hormozdiari et al. 2009; Campbell et al. 2008) and Third Generation Technologies (Maron et al. 2013) were exploited. Sanger sequences are about 1 kb long with nearly perfect accuracy and can be produced only at very low throughput and high costs. Second Generation sequences, e.g. Illumina sequences, are much shorter, 100–150 bp for HiSeq machines and 250–300 bp for MiSeq instruments, of good accuracy with only 1 % erroneous bases and throughput increase of orders of magnitude with several Gb produced daily at very limited cost. Finally, Third Generation, single molecule-derived, sequences, e.g. PacBio sequences, are a few kb long, still of limited accuracy with more than 10 % erroneous bases, but with dozens of Mb produced daily at limited costs. Therefore, apart from timing and costs, Second and Third Generation sequences mainly differ on read length and accuracy, and throughput. These factors highly influence the kind of methods to be used to tackle the problem of CNV detection.

NGS Technologies vs. Computational Techniques

Most of the current algorithms for SVs detection are modeled on computational methods that were initially developed to analyze Sanger sequences (Tuzun et al. 2005; Korbel et al. 2007; Kidd et al. 2008; Korbel et al. 2009). So far, NGS based methods to detect SVs can be categorized into five different strategies: paired-end mapping (PEM), split read mapping (SRM), depth of coverage (DOC), de novo assembly (DNA), and a combination of the above approaches (COMBI). PEM was historically the first method based on sequencing used to discover SVs (Tuzun et al. 2005). It assesses the span and orientation of paired reads detecting discordant pairs whose orientation is not as expected or distance is significantly different from the predetermined average insert size. PEM-based tools applied to NGS datasets usually search for clusters of such signatures (Medvedev et al. 2009). Multiple evidences are required to strengthen the signal of usually short NGS reads. Gathering information from multiple evidences is called clustering and it can be divided into hard and soft clustering. In hard clustering reads that map to multiple locations are discarded in order to avoid false positive SVs due to repetitive regions. In soft clustering (Hormozdiari et al. 2009), instead, in order to improve sensitivity, such reads are not discarded and assigned to a single cluster. PEM-based tools can be used to detect effectively deletions, short insertions, inversions, translocations, and

duplications at almost single base pair resolution. Insertions size is limited to library insert size unless multiple evidences are used, e.g. clusters of single reads with only one read of the pair of a fragment can safely be positioned in the genome may hint the insertion of a repetitive element (Platzer et al. 2012; Fiston-Lavier et al. 2011). PEM methods based either on hard or soft clustering suffer, respectively, of sensitivity and specificity in repetitive and low-complexity regions. SRM methods are based on gapped alignment of a single read to the reference genome and can be used to determine SV breakpoints down to base pair resolution. If a read does not align entirely, then a gapped alignment is applied.

SRM was first applied to long Sanger reads (Mills et al. 2006) but later SRM methods were developed also for the NGS technologies with some modifications: (1) given the high coverage of NGS experiments, clusters of split reads are requested as proper signature, (2) given the short length of NGS sequences, split reads usually tend to map to multiple locations of a genome. To overcome this problem the mapping of their mate is used as a reliable anchor, severely limiting the search space for the split read, (3) elongate single reads producing overlapping paired read libraries that can be merged into a single longer read, (4) complement PEM methods providing putative SVs with breakpoints not determined at base pair resolution. In general, SRM methods heavily rely on the length of reads, still a problem for Second Generation Sequences, and are not applicable to repetitive or low-complexity regions. SRM-based tools can be used to detect effectively deletions and very short insertions at base pair resolution. The limitation on insertions is given by the read length itself. DOC methods are still based on read alignment but unlike PEM and SRM methods, they mainly care on DOC and less on single base resolution. Their main assumption is that the number of mapping reads follows a Poisson distribution and regions deleted or duplicated will have less or more reads assigned to them, respectively. DOC-based tools can be classified in at least two categories: single sample and multi-sample. In the first case, average read depth is estimated using mathematical models and then regions that depart from it are discovered. In the second case, a sample is used as control and all other samples are compared to its coverage rather than to an average read depth. Therefore, in the first case copy numbers are absolute numbers while in the second case they are relative to the control sample. In general, DOC methods follow a four-step procedure composed of: (1) independently mapping reads of each sample towards a reference genome, (2) normalize coverage along a sliding window where read depth of a single window is computed according to the number of reads mapped in it (normalization basically serves to correct potential biases in read depths mainly caused by GC content and repetitive regions), (3) estimation of copy number, either absolute or relative, along the sliding window in order to determine possible gain or loss, (4) segmentation, merging adjacent genomic regions with a similar copy number using statistical models. Sliding windows can be computed in a variety of ways as with a fixed width or with a predefined amount of reads mapping within it. DOC methods can detect CNVs with respect to what is present in the reference genome, therefore novel insertions and inversions cannot be detected. The detection reveals the copy number but not the location of possible new copies. The breakpoint resolution is very poor and is on the level of several kb.

Methods based on DNA differ from previously described methods as they do not rely on a first step of read alignment toward a reference genome but directly use the reads to assemble them into contigs that are later compared to a reference genome in order to discover discrepancies. Comparison is usually performed through sequence alignment to the reference genome. An alternative approach is proposed by the software Cortex (Iqbal et al. 2012; Leggett et al. 2013), designed to directly discover CNVs among multi-samples: as most assemblers it is based on de Bruijn graphs with the exception that nodes and edges are marked in different colors to differentiate different samples. Unfortunately, although a range of assemblers have been developed (Simpson et al. 2009; Gnerre et al. 2011; Luo et al. 2012; Simpson and Durbin 2012; Zimin et al. 2013), given the short length of NGS sequences, DNA is still challenging and the accuracy of contigs produced is unsatisfactory especially in repetitive regions that are often a great source of variations. An emerging branch in the field is the assembly of limited regions, e.g. exomes or fosmid clones, which should lead to improved assemblies and consequently improved CNV detection, though at the cost of restricting to limited portions of the genome.

Although methods based on the four previously described categories (PEM, SRM, DOC, and DNA) have been greatly improved and a wide number of tools have been recently developed, none is able to reliably detect SVs, either in terms of sensitivity or specificity. Each has different strengths and weaknesses in detection, depending on the kind of variant or the sequence at the studied *locus*. To overcome the implicit limitations of individual methods, often operating in a complementary manner, it is possible to implement approaches (COMBI) including the different methods and therefore improve the detection performance and reduce the number of false positives. While PEM and SRM methods are related to each other, the other methods, DOC and DNA represent complementary methodologies that could benefit from each other. In some cases, they can detect identical events, perhaps with different strength and precision, while in other cases they can detect very independent events that cannot be discovered by all methodologies.

Future Perspectives

Examples of association of SVs to agronomically relevant phenotypes can be found in a recent review on putative dispensable regions of plant genomes (Marroni et al. 2014). The repertoire of functional and evolutionary consequences of SVs is expanding, but a comprehensive map of all causative SVs is still far from complete. The advent of NGS technologies highly improved the detection rate of SVs even if such technologies are affected by two main drawbacks: difficulty with reliably mapping short reads to DNA repeats (Treangen and Salzberg 2012) and platform-specific biases, which result in lower read coverage of some parts of the genome (for example, GC-rich regions) (Dohm et al. 2008). Compared to detection of SVs using a single tool, the combination of different software has proved effective in overcoming the main drawbacks of NGS technologies and in improving SVs prediction accuracy. In addition, the use of libraries with different characteristics has been

proved effective in the detection of SVs. Moreover, longer reads may greatly improve the specificity of reads mapping and consequently SVs detection. In this context, third generation sequencing (TGS) provide reads as long as few kb and could solve most of the problems of shorter reads, in particular in presence of repetitive regions source of most misalignments. Currently, its main problem is the accuracy of base calls, much lower than most of previous sequencing technologies. A final comment on the availability of proper infrastructures for SVs detection is needed. The huge quantity of NGS data requires a large hardware infrastructure to handle it in terms of both disk space and computational resources. Comprehensive databases of already discovered SVs could highly improve the detection but also the evaluation of putative newly discovered SVs. Although already available for human genetics the Database of Genomic Variants (http://dgv.tcag.ca/dgv/app/home; http://www.ncbi.nlm.nih.gov/dbvar/; http://www.ebi.ac.uk/dgva/), the need for an SV database for the plant kingdom could be seriously considered by bioinformatics institutes in the near future.

On the applicative, plant breeding side, we should consider whether and how CNVs will be effectively used for genomic-assisted selection. A relevant starting consideration is that for too much time this kind of variation has been excluded from genetic association studies. Not only in plants, but also in humans, sometimes genetic risk factors have been missed because association studies have sought risk-associated SNPs, while ignoring structural variation causing gene copy number changes, as reported for colorectal adenoma recurrence (Laukaitis et al. 2010). To avoid this, as anticipated by Stranger et al. (2007) and by Beckmann et al. (2007) for humans, the interrogation of the genomes for both types of variants (SNPs and CNVs) in association studies may be an effective way to elucidate the causes of complex phenotypes in humans, animals, and plants.

Acknowledgements This work was partially supported by the FROSTMAP project of the Fondazione Cassa di Risparmio di Modena and by IGA Technology Services.

References

Abel HJ, Duncavage EJ (2013) Detection of structural DNA variation from next generation sequencing data: a review of informatic approaches. Cancer Genet 206(12):432–440

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363–376

Bailey JA, Kidd JM, Eichler EE (2008) Human copy number polymorphic genes. Cytogenet Genome Res 123(1–4):234–243

Beckmann JS, Estivill X, Antonarakis SE (2007) Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. Nat Rev Genet 8(8):639–646

Beló A, Beatty MK, Hondred D, Fengler KA, Li B, Rafalski A (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. Theor Appl Genet 120(2):355–367

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59

Bickhart DM, Hou Y, Schroeder SG, Alkan C, Cardone MF, Matukumalli LK et al (2012) Copy number variation of individual cattle genomes using next-generation sequencing. Genome Res 22(4):778–790

Birchler JA (2012) Insights from paleogenomic and population studies into the consequences of dosage sensitive gene expression in plants. Curr Opin Plant Biol 15(5):544–548

Boyko A, Kovalchuk I (2011) Genetic and epigenetic effects of plant-pathogen interactions: an evolutionary perspective. Mol Plant 4(6):1014–1023

Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 40(6):722–729

Cantsilieris S, White SJ (2013) Correlating multiallelic copy number polymorphisms with disease susceptibility. Hum Mutat 34(1):1–13

Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J et al (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43(10):956–963

Castonguay Y, Dubé M-P, Cloutier J, Bertrand A, Michaud R, Laberge S (2013) Molecular physiology and breeding at the crossroads of cold hardiness improvement. Physiol Plant 147(1): 64–74

Causse M, Desplat N, Pascual L, Le Paslier M-C, Sauvage C, Bauchet G et al (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. BMC Genomics 14:791

Chaignat E, Yahya-Graison EA, Henrichsen CN, Chrast J, Schütz F, Pradervand S et al (2011) Copy number variation modifies expression time courses. Genome Res 21(1):106–113

Chia J-M, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J et al (2012) Maize HapMap2 identifies extant variation from a genome in flux. Nat Genet 44(7):803–807

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y et al (2010) Origins and functional impact of copy number variation in the human genome. Nature 464(7289):704–712

Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM et al (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. Science 338(6111): 1206–1209

D'haene B, Vandesompele J, Hellemans J (2010) Accurate and objective copy number profiling using real-time quantitative PCR. Methods 50(4):262–270

Daines B, Wang H, Li Y, Han Y, Gibbs R, Chen R (2009) High-throughput multiplex sequencing to discover copy number variants in Drosophila. Genetics 182(4):935–941

DeBolt S (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. Genome Biol Evol 2:441–453

Díaz A, Zikhali M, Turner AS, Isaac P, Laurie DA (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (*Triticum aestivum*). PLoS One 7(3):e33234

Diskin SJ, Hou C, Glessner JT, Attiyeh EF, Laudenslager M, Bosse K et al (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459(7249):987–991

Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 36(16):e105

Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7(2):85–97

Fiston-Lavier A-S, Carrigan M, Petrov DA, González J (2011) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Res 39(6):e36

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM et al (2006) Copy number variation: new insights in genome diversity. Genome Res 16(8):949–961

Girirajan S, Campbell CD, Eichler EE (2011) Human copy number variation and complex genetic disease. Annu Rev Genet 45:203–226

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 108(4):1513–1518

Golzio C, Willer J, Talkowski ME, Oh EC, Taniguchi Y, Jacquemont S et al (2012) KCTD13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant. Nature 485(7398):363–367

Haber JE (2000) Partners and pathways repairing a double-strand break. Trends Genet 16(6): 259–264

Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009a) Mechanisms of change in gene copy number. Nat Rev Genet 10(8):551–564

Hastings PJ, Ira G, Lupski JR (2009b) A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5(1):e1000327

Haun WJ, Hyten DL, Xu WW, Gerhardt DJ, Albert TJ, Richmond T et al (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar Williams 82. Plant Physiol 155(2):645–655

Henrichsen CN, Vinckenbosch N, Zöllner S, Chaignat E, Pradervand S, Schütz F et al (2009) Segmental copy number variation shapes tissue transcriptomes. Nat Genet 41(4):424–429

Hiroi N, Takahashi T, Hishimoto A, Izumi T, Boku S, Hiramoto T (2013) Copy number variation at 22q11.2: from rare variants to common mechanisms of developmental neuropsychiatric disorders. Mol Psychiatry 18(11):1153–1165

Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Res 19(7):1270–1278

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y et al (2004) Detection of large-scale variation in the human genome. Nat Genet 36(9):949–951

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet 44(2):226–232

Jacquemont S, Reymond A, Zufferey F, Harewood L, Walters RG, Kutalik Z et al (2011) Mirror extreme BMI phenotypes associated with gene dosage at the chromosome 16p11.2 locus. Nature 478(7367):97–102

Jiao Y, Zhao H, Ren L, Song W, Zeng B, Guo J et al (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44(7):812–815

Kellogg EA, Bennetzen JL (2004) The evolution of nuclear genome structure in seed plants. Am J Bot 91(10):1709–1725

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56–64

Knox AK, Dhillon T, Cheng H, Tondelli A, Pecchioni N, Stockinger EJ (2010) CBF gene copy number variation at Frost Resistance-2 is associated with levels of freezing tolerance in temperate-climate cereals. Theor Appl Genet 121(1):21–35

Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc Biol Sci 279(1749):5048–5057

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF et al (2007) Paired-end mapping reveals extensive structural variation in the human genome. Science 318(5849):420–426

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z et al (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10(2):R23

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H et al (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42(11):1027–1030

Laukaitis CM, Thompson P, Martinez ME, Gerner EW (2010) Identifying gene copy number variants associated with colorectal adenoma recurrence. Genome Biol 11(Suppl 1):24

Lee JA, Carvalho CMB, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. Cell 131(7):1235–1247

Leggett RM, Ramirez-Gonzalez RH, Verweij W, Kawashima CG, Iqbal Z, Jones JDG et al (2013) Identifying and classifying trait linked polymorphisms in non-reference species by walking coloured de Bruijn graphs. PLoS One 8(3):e60058

Li W, Olivier M (2013) Current analysis platforms and methods for detecting copy number variation. Physiol Genomics 45(1):1–16

Lieberman-Lazarovich M, Levy AA (2011) Homologous recombination in plants: an antireview. Methods Mol Biol 701:51–65
Lower KM, Hughes JR, De Gobbi M, Henderson S, Viprakasit V, Fisher C et al (2009) Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proc Natl Acad Sci U S A 106(51):21771–21776
Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):18
Lupski JR (2007) Genomic rearrangements and sporadic disease. Nat Genet 39(7 Suppl): S43–S47
Lupski JR, de Oca-Luna RM, Slaugenhaupt S, Pentao L, Guzzetta V, Trask BJ et al (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66(2):219–232
Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci U S A 107(20):9270–9274
Maron LG, Guimarães CT, Kirst M, Albert PS, Birchler JA, Bradbury PJ et al (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. Proc Natl Acad Sci U S A 110(13):5241–5246
Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18C:31–36
McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A et al (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 40(10):1166–1174
McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL et al (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159(4):1295–1308
Medvedev P, Stanciu M, Brudno M (2009) Computational methods for discovering structural variation with next-generation sequencing. Nat Methods 6(11 Suppl):S13–S20
Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS et al (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16(9): 1182–1190
Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C et al (2011) Mapping copy number variation by population-scale genome sequencing. Nature 470(7332):59–65
Morel M, Meux E, Mathieu Y, Thuillier A, Chibani K, Harvengt L et al (2013) Xenomic networks variability and adaptation traits in wood decaying fungi. Microb Biotechnol 6(3):248–263
Muñoz-Amatriaín M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B et al (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. Genome Biol 14(6):R58
Nitcher R, Distelfeld A, Tan C, Yan L, Dubcovsky J (2013) Increased copy number at the HvFT1 locus is associated with accelerated flowering time in barley. Mol Genet Genomics 288(5–6):261–275
Oh D-H, Dassanayake M, Bohnert HJ, Cheeseman JM (2012) Life at the extreme: lessons from the genome. Genome Biol 13(3):241
Pâques F, Haber JE (1999) Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiol Mol Biol Rev 63(2):349–404
Park H, Kim J-I, Ju YS, Gokcumen O, Mills RE, Kim S et al (2010) Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. Nat Genet 42(5):400–405
Perry GH, Tchinda J, McGrath SD, Zhang J, Picker SR, Cáceres AM et al (2006) Hotspots for copy number variation in chimpanzees and humans. Proc Natl Acad Sci U S A 103(21): 8006–8011
Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R et al (2007) Diet and the evolution of human amylase gene copy number variation. Nat Genet 39(10):1256–1260
Platzer A, Nizhynska V, Long Q (2012) TE-Locate: a tool to locate and group transposable element occurrences using paired-end next-generation sequencing data. Biology 1(2):395–410

Puchta H (2005) The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. J Exp Bot 56(409):1–14

Puchta H, Hohn B (1991) A transient assay in plant cells reveals a positive correlation between extrachromosomal recombination rates and length of homologous overlap. Nucleic Acids Res 19(10):2693–2700

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD et al (2006) Global variation in copy number in the human genome. Nature 444(7118):444–454

Saintenac C, Jiang D, Akhunov ED (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. Genome Biol 12(9):R88

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP et al (2007) Challenges and standards in integrating surveys of structural variation. Nat Genet 39(7 Suppl):S7–S15

Schnable PS, Springer NM (2013) Progress toward understanding heterosis in crop plants. Annu Rev Plant Biol 64:71–88

Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. Proc R Soc B Biol Sci 277(1698):3213–3221

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P et al (2004) Large-scale copy number polymorphism in the human genome. Science 305(5683):525–528

Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. Genome Res 22(3):549–556

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia 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

Stambuk BU, Dunn B, Alves SL Jr, Duval EH, Sherlock G (2009) Industrial fuel ethanol yeasts contain adaptive copy number changes in genes involved in vitamin B1 and B6 biosynthesis. Genome Res 19(12):2271–2278

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813): 848–853

Sutton T, Baumann U, Hayes J, Collins NC, Shi B-J, Schnurbusch T et al (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. Science 318(5855): 1446–1449

Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D et al (2010) Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res 20(12):1689–1699

Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13(1):36–46

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM et al (2005) Fine-scale structural variation of the human genome. Nat Genet 37(7):727–732

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30(1):105–111

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res 19(9):1586–1592

Yu P, Wang C-H, Xu Q, Feng Y, Yuan X-P, Yu H-Y et al (2013) Genome-wide copy number variations in *Oryza sativa* L. BMC Genomics 14:649

Zhang J, Zuo T, Peterson T (2013) Generation of tandem direct duplications by reversed-ends transposition of maize ac elements. PLoS Genet 9(8):e1003691

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S et al (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (Sorghum bicolor). Genome Biol 12(11):R114

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29(21):2669–2677

Zmieńko A, Samelak A, Kozłowski P, Figlerowicz M (2014) Copy number polymorphism in plant genomes. Theor Appl Genet 127:1–18

Index

G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9