Methods in Molecular Biology 1667

Springer Protocols

Gaurav Sablok
Hikmet Budak
Peter J. Ralph *Editors*

Brachypodium Genomics

Methods and Protocols

EXTRAS ONLINE

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Brachypodium Genomics

Methods and Protocols

Edited by

Gaurav Sablok

Department of Biodiversity and Molecular Ecology, Fondazione Edmund Mach, IASMA, Italy

Hikmet Budak

Sabanci University, Istanbul, Turkey

Peter J. Ralph

Climate Change Cluster (C3), University of Technology Sydney, Sydney, NSW, Australia

Humana Press

Editors
Gaurav Sablok
Department of Biodiversity
and Molecular Ecology
Fondazione Edmund Mach
IASMA, Italy

Hikmet Budak
Sabanci University
Istanbul, Turkey

Peter J. Ralph
Climate Change Cluster (C3)
University of Technology Sydney
Sydney, NSW, Australia

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-7276-0 ISBN 978-1-4939-7278-4 (eBook)
DOI 10.1007/978-1-4939-7278-4

Library of Congress Control Number: 2017952741

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

Model plant genomics has been revolutionized after the advent of the next-generation sequencing technologies, with major focus being leveraged on the application of these techniques to understand the functional 'ome of these species. Recalling the publication of the first flowering plant genome (*Arabidopsis thaliana*) in 2000 (Arabidopsis Genome Initiative, 2000), achievable and demonstrated advancements in model plant genomics have been achieved and a member of Pooideae clade, *Brachypodium distachyon*, was proposed and sequenced as a monocot model for comparative plant genomics. Since the first reports on the genome of *Brachypodium distachyon*, significant progress and swathing amount of information and application of trait genomics in this model species has revealed several key concepts for advancing monocot genomics and also laid the foundation for understanding the complex transcriptional machinery and regulatory mechanism and initiating the cross talks between imprinting and cross-linking the genes with epigenetic variations. This volume enlists a comprehensive layout of protocols for *Brachypodium* genomics in several domains ranging from marker development, trait evolution, functional genomics, metabolomics, transcriptomics, and genomics to tilling-based approaches, which will play a key role in advancing the *Brachypodium* genomics. This volume will not only play a key role in providing the standard protocols for widening the genetic base of *Brachypodium*, but will also help elucidate the development and advances in understanding the model plant in question using NGS technologies. This volume bridges the gap between the bench-oriented molecular biologist and computational biologist working toward accelerated and evolving *Brachypodium* genomics. I would like to thank Associate Professor Xiang Jia Jack Min, Department of Biological Sciences, Youngstown State University, Youngstown, OH 44555, USA, for support on *Brachypodium* and Dr. Heidi Hauffe, Head of Department of Biodiversity and Molecular Ecology, Fondazione Edmund Mach, IASMA, 380010, Italy, for encouraging me to conceive and design this proposed volume. This volume would haven't been possible without the support of Namrata Sablok, who patiently supported me during the intense work on this volume, and also I would like to thank my family for their support during the final preparation of the volume.

IASMA, Italy ***Gaurav Sablok***

Contents

Contributors

Elizabeth A.B. Aitken • *School of Agriculture and Food Science, The University of Queensland, Brisbane, QLD, Australia*

Louise de Bang • *Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg, Denmark*

Renesh H. Bedre • *Texas A&M AgriLife Research & Extension Center, Weslaco, TX, USA*

Edoardo Bertolini • *Donald Danforth Plant Science Center, St. Louis, MO, USA; Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy*

Alexander Betekhtin • *Department of Plant Anatomy and Cytology, Faculty of Biology and Environmental Protection, University of Silesia in Katowice, Katowice, Poland*

Hikmet Budak • *Faculty of Engineering and Natural Sciences, Molecular Biology, Genetics and Bioengineering Program, Sabanci University, Istanbul, Turkey; Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, MT, USA*

Salvatore Camiolo • *Dipartimento di Agraria, SACEG, Università degli studi di Sassari, Sassari, Italy*

Kranthi Chennamsetti • *Centre for Bioinformatics Research (CBR), SRKR Engineering College, Bhimavaram, Andhra Pradesh, India*

Timothy L. Fitzgerald • *Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture, Brisbane, QLD, Australia*

Upinder S. Gill • *Noble Research Institute, LLC., Ardmore, OK, USA*

Korneliya Gudys • *Department of Genetics, University of Silesia in Katowice, Katowice, Poland*

Robert J. Henry • *Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD, Australia*

Dominika Idziak-Helmcke • *Department of Plant Anatomy and Cytology, Faculty of Biology and Environmental Protection, University of Silesia in Katowice, Katowice, Poland*

Sonia Irigoyen • *Texas A&M AgriLife Research & Extension Center, Weslaco, TX, USA*

Aneesh Karunakaran • *Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture, Brisbane, QLD, Australia; School of Agriculture and Food Science, The University of Queensland, Brisbane, QLD, Australia*

Kemal Kazan • *Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture, Brisbane, QLD, Australia; Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD, Australia*

Can Kou • *Department of Life Science and Technology, Lijiang Teachers College, Lijiang City, Yunnan, People's Republic of China*

Miroslaw Kwasniewski • *Department of Genetics, University of Silesia in Katowice, Katowice, Poland*

Hee-Kyung Lee • *Noble Research Institute, LLC., Ardmore, OK, USA*

I-Hsuan Lin • *VGH-YM Genome Center, National Yang-Ming University, Taipei, Taiwan*

Lifeng Liu • *Energy Bioscience Institute, University of California, Berkeley, CA, USA*

Kranthi K. Mandadi • *Texas A&M AgriLife Research & Extension Center, Weslaco, TX, USA; Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, USA*

Erica Mica • *Genomics Research Centre, Consiglio per la Ricerca in Agricoltura e L'analisi Dell'Economia Agraria, Fiorenzuola d'Arda, Italy*
Xin-Chun Mo • *Department of Life Science and Technology, Lijiang Teachers College, Lijiang City, Yunnan, People's Republic of China*
Keiichi Mochida • *Cellulose Production Research Team, RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan; Kihara Institute for Biological Research, Yokohama City University, Yokohama, Kanagawa, Japan*
Suresh B. Mudunuri • *Centre for Bioinformatics Research (CBR), SRKR Engineering College, Bhimavaram, Andhra Pradesh, India*
Kirankumar S. Mysore • *Noble Research Institute, LLC., Ardmore, OK, USA*
Bahar Sogutmaz Ozdemir • *Faculty of Engineering, Department of Genetics and Bioengineering, Yeditepe University, Istanbul, Turkey*
Mario Enrico Pè • *Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy*
Nita Parekh • *International Institute of Information Technology, Hyderabad, Telangana, India*
Andrea Porceddu • *Dipartimento di Agraria, SACEG, Università degli studi di Sassari, Sassari, Italy*
Jonathan J. Powell • *Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture, Brisbane, QLD, Australia; Queensland Alliance for Agriculture and Food Innovation, The University of Queensland, Brisbane, QLD, Australia*
Concetta De Quattro • *Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy*
Anuj Rana • *Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture, Brisbane, QLD, Australia; Department of Genetics, University of Delhi, New Delhi, India*
Søren K. Rasmussen • *Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg, Denmark*
Rosalie Sabburg • *Commonwealth Scientific and Industrial Research Organisation (CSIRO) Agriculture, Brisbane, QLD, Australia*
Gaurav Sablok • *Department of Biodiversity and Molecular Ecology, Fondazione Edmund Mach, IASMA, Italy; Plant Functional Biology and Climate Change Cluster (C3), University of Technology Sydney, Sydney, NSW, Australia*
Karen-Beth G. Scholthof • *Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, USA*
Juan C. Serrani-Yarce • *Noble Research Institute, LLC., Ardmore, OK, USA; Department of Biological Sciences, University of North Texas, Denton, TX, USA*
Kazuo Shinozaki • *Gene Discovery Research Group, RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan; Biomass Research Platform Team, RIKEN Center for Sustainable Resource Science, Yokohama, Kanagawa, Japan*
Sanchari Sircar • *International Institute of Information Technology, Hyderabad, Telangana, India*
Anna Maria Torp • *Department of Plant and Environmental Sciences, University of Copenhagen, Frederiksberg, Denmark*
G.P. Saradhi Varma • *Centre for Bioinformatics Research (CBR), SRKR Engineering College, Bhimavaram, Andhra Pradesh, India*
Ron Vunsh • *Department of Plant and Environmental Sciences, The Weizmann Institute of Science, Rehovot, Israel*

Xiaopeng Wen • *Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region (Guizhou University), Ministry of Education, Institute of Agro-bioengineering, Guizhou University, Guiyang, Guizhou, People's Republic of China; College of Life sciences, Guizhou University, Guiyang, Guizhou, People's Republic of China*

Kun Yang • *Key Laboratory of Plant Resources Conservation and Germplasm Innovation in Mountainous Region (Guizhou University), Ministry of Education, Institute of Agro-bioengineering, Guizhou University, Guiyang, Guizhou, People's Republic of China; College of Life sciences, Guizhou University, Guiyang, Guizhou, People's Republic of China*

Ling-Juan Yin • *National Technical Secondary School of Lijiang, Lijiang, Yunnan, People's Republic of China*

De-Quan Zhang • *College of Pharmacy and Chemistry, Dali University, Dali, Yunnan, People's Republic of China*

Zhenguo Zhang • *Department of Biology, University of Rochester, Rochester, NY, USA*

Chapter 1

Methods for Cytogenetic Chromosome Barcoding and Chromosome Painting in *Brachypodium distachyon* and Its Relative Species

Dominika Idziak-Helmcke and Alexander Betekhtin

Abstract

Brachypodium distachyon provides a particularly appealing object for molecular cytogenetic analysis due to its compact genome and low repetitive DNA content, as well as low ($x = 5$) basic number of chromosomes easily identifiable on the basis of their morphometric features. Some of these features, such as genome compactness, are shared by the other members of the genus, thus making them amenable for comparative cytogenetic mapping. Cytogenetic infrastructure established for *B. distachyon* was initially based on fluorescence in situ hybridization with various tandemly repeated sequences as probes. The molecular cytogenetic studies advanced greatly with the development of *B. distachyon* large DNA insert genomic libraries. These resources coupled with the access to the fully sequenced genome of *B. distachyon* enabled chromosome painting in monocots for the first time. This pioneering work was subsequently extended to other *Brachypodium* species, allowing insight into grass karyotype evolution. In this protocol we describe the methods of making somatic and meiotic chromosome preparations, probe labeling, FISH with BAC clones, a strategy for chromosome barcoding and chromosome painting in *B. distachyon*, and comparative chromosome painting in the other *Brachypodium* species.

Key words Bacterial artificial chromosome (BAC), *Brachypodium*, Chromosome preparation, Chromosome barcoding, Chromosome painting, Fluorescence in situ hybridization (FISH), Meiosis, Mitosis, Molecular cytogenetics

1 Introduction

Fluorescence in situ hybridization (FISH) is one of the most utilized molecular cytogenetic methods that allows for studying genome structure and evolution at the microscopic level. It is based on kinetically controlled annealing of fluorescently labeled DNA or RNA molecules (probes) to the complementary sequences in the cytological or histological preparations and subsequent detection of such probes by fluorescent microscopy. The DNA: DNA FISH enables direct visualization of the targeted loci in the mitotic and meiotic chromosomes as well as interphase nuclei, thus

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_1,

providing insight into various aspects of genome organization and dynamics.

Brachypodium distachyon, a recently adopted model for temperate cereals and grasses, provides a valuable object for cytomolecular analyses due to its desirable biological features such as compact genome of ca. 320 Mb, low basic chromosome number ($x = 5$), large and conspicuous chromosomes, and asymmetric karyotype [1, 2]. Some of these attributes are shared by the other members of the genus *Brachypodium*. The genus comprises less than 20 species that display a significant degree of diversity in basic chromosome number and ploidy level, as well as differences in chromosome size and morphology. Due to these features, *Brachypodium* represents a particularly suitable model system for studying the evolution and divergence of grass karyotypes.

A well-developed set of experimental tools including large collections of accessions, fully sequenced nuclear genome, and availability of ordered BAC libraries of genomic DNA [3–5] has led to rapid advances in establishing the cytomolecular platform for *B. distachyon* and, in turn, to the analysis of its genome by FISH-based techniques with unprecedented precision [6]. Moreover, the feasibility of comparative FISH-based mapping of the karyotypes of various *Brachypodium* species using the *B. distachyon* resources has been demonstrated numerous times [7–9]. The results of comparative studies allowed the inference of the mechanisms responsible in particular for shaping the karyotype evolution in the genus and in general for the divergence of grass genomes [10].

In this chapter we present two approaches to the cytogenetic analysis of karyotype structure and evolution in *B. distachyon* and other members of the genus: chromosome painting (CP) and chromosome barcoding (CB). CP is possibly the most advanced and powerful technique used in modern cytogenetics. It is based on using large pools of chromosome-specific, low-repeat BAC clones as FISH probes. CP enables the visualization and tracking of individual chromosomes or their segments at various stages of mitotic and meiotic cell division (Fig. 1a). Its variant, comparative chromosome painting (CCP), can be used to identify homeologous chromosomes in the karyotypes of closely related species and to detect nested chromosome fusions and other large-scale chromosome rearrangements that are responsible for the karyotype differentiation (Fig. 1b). However, an apparent limitation of CP and CCP results from using probes that span a significant part of the whole arm or even the entire chromosome, which may preclude the detection of small-scale reshuffling of the chromosome structure. This problem can be solved by using CB, which utilizes single BAC clones as probes in multicolor FISH experiments (Fig. 1c, d). CB represents a complementary approach to CP since it allows studying the karyotype structure in more detail and the making of high-density cytogenetic maps. A comparative

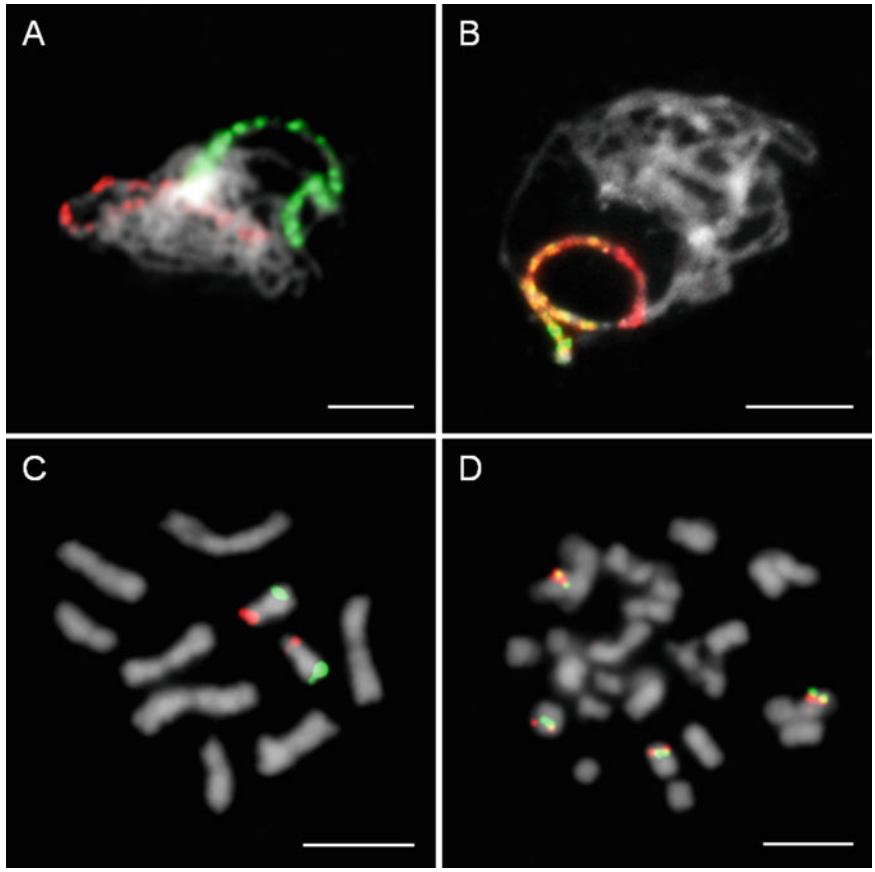

Fig. 1 Cytogenetic BAC-FISH mapping the chromosomes of various *Brachypodium* species. (**a**) CP in pachytene bivalents of *B. distachyon* using BAC pools spanning the short (*green* fluorescence) and long (*red* fluorescence) arm of chromosome Bd2. (**b**) CCP in pachytene bivalents of *B. pinnatum* ($2n = 16$) using BAC pools spanning the short (*green* fluorescence) and long (red fluorescence) arm of chromosome Bd5. (**c**) CB in metaphase chromosomes of *B. distachyon* ($2n = 10$) using clones BD_CBa0038G13 (*green* fluorescence) and BD_ABa0040K04 (*red* fluorescence). (**d**) CCB in metaphase chromosomes of *B. phoenicoides* ($2n = 28$) using clones BD_ABa0044B16 (*green* fluorescence) and BD_ABa0005E09 (*red* fluorescence). Chromosomes were counterstained with DAPI. Bar: 5 mm

chromosome barcoding (CCB) helps to uncover karyotypical changes with higher resolution and facilitates studies of chromosomal evolutionary processes.

2 Materials

2.1 Mitotic and Meiotic Chromosome Preparations

1. Carnoy's fixative: methanol and glacial acetic acid (3:1), freshly prepared (*see* **Note 1**).
2. 0.1 M citrate buffer: Prepare 0.1 M citric acid monohydrate (solution A, 21.01 g/L) and 0.1 M trisodium citrate dihydrate (solution B, 28.41 g/L) in distilled water. Prepare the stock

buffer solution by mixing A and B at a ratio of 4:6 and adjust pH to 4.8 using A or B. Store the stock solution at −20 °C.

3. 0.01 M citrate buffer: dilute 0.1 M citrate buffer with distilled water. Store the buffer solution at 4 °C.
4. Enzyme mixture for digesting root meristems: 4% (vol/vol) pectinase (Sigma), 1% (wt/vol) cellulase (Calbiochem), 1% (wt/vol) cellulase "Onozuka" R-10 (Serva). Dissolve all reagents in 0.01 M citrate buffer by stirring on a magnetic stirrer. Divide the enzyme into 1 mL aliquots and store at −20 °C. Enzyme can be reused several times.
5. Enzyme mixture for digesting anthers: 10% (vol/vol) pectinase (Sigma), 0.65% (wt/vol) cellulase "Onozuka" R-10 (Serva), 0.5% (wt/vol) cellulase (Calbiochem), 0.15% (wt/vol) cytohelicase (Sigma), and 0.15% (wt/vol) pectolyase (Sigma). Dissolve all reagents in 0.01 M citrate buffer by stirring on a magnetic stirrer. Divide the enzyme into 1 mL aliquots and store at −20 °C. Enzyme can be reused several times.
6. 45% acetic acid.

2.2 BAC DNA Isolation

1. LB Agar: Prepare 1 L LB Agar in distilled water. Sterilize by autoclaving.
2. LB Broth: Prepare 1 L LB Broth in distilled water. Sterilize by autoclaving.
3. 25 mg/mL chloramphenicol in isopropanol: Add to the LB Agar and LB Broth until the final concentration of chloramphenicol equals 12.5 μg/mL (*see* **Note 2**).
4. Solution I (*see* **Note 3**): 50 mM Tris–HCl, pH 8.0, 10 mM EDTA, 100 μg/mL RNase A. Mix 10 mL of 0.1 M EDTA and 10 mL of 0.5 M Tris–HCl and bring the volume to 100 mL with sterile distilled water (SDW). Dilute 10 mg/mL RNase A stock in the prepared solution in the proportion of 1:99. RNase A should be freshly added to the solution just before use.
5. 0.1 M EDTA: Add 3.72 g of Na_2EDTA to 70 mL of distilled water. While stirring vigorously on a magnetic stirrer, add NaOH pellet or 10 N NaOH to adjust pH to 8.0. Bring the volume to 100 mL with distilled water. Sterilize by autoclaving and store at room temperature (RT).
6. 0.5 M Tris–HCl: Add 6.057 g of Trizma Base to 70 mL of distilled water and adjust pH to 8.0 with HCl. Bring the volume to 100 mL with distilled water. Sterilise by autoclaving and store at 4 °C.
7. 10 mg/mL RNase A: Dissolve 100 mg RNase A in 10 mL of 10 mM Tris–HCl + 15 mM NaCl. Boil for 15 min and allow to cool. Store in 1 mL aliquots at −20 °C.

8. 10 mM Tris–HCl, pH 8.5: Add 1.211 g of Trizma Base to 700 mL of distilled water. Adjust to pH 8.5 with HCl. Bring the volume to 1 L with distilled water. Sterilize by autoclaving and store at RT.
9. Solution II (*see* **Note 3**): 0.2 M NaOH, 1% sodium dodecyl sulfate (SDS). Mix 10 mL of 10% SDS and 20 mL of 1 M NaOH and bring the volume to 100 mL with SDW.
10. 1 M NaOH: Dissolve 4 g of NaOH in 100 mL of SDW.
11. 10% sodium dodecyl sulfate (SDS): Dissolve 10 g of SDS in 80 mL of SDW and then add SDW to 100 mL. Store at RT.
12. Solution III (*see* **Note 3**): 3 M KOAc pH 5.5. Dissolve 54 g of KOAc in 80 mL of distilled water. Adjust pH to 5.5 and then add distilled water to 100 mL. Sterilize by autoclaving and store at RT.
13. EB Buffer: 10 mM Tris–HCl, pH 8.5.

2.3 Probe Labeling

2.3.1 Probe Labeling Using Commercially Available Labeled Nucleotides

1. 0.4 mM dATP, 0.4 mM dCTP, 0.4 mM dGTP, and 0.4 mM dTTP: Keep 25 µl aliquots of each dNTP at −20 °C.
2. Tetramethylrhodamine-5-dUTP (Roche) or digoxigenin-11-dUTP (Roche).
3. Nick Translation Mix (Roche).

2.3.2 Probe Labeling Using Custom-Made Labeled Nucleotides

1. DMSO.
2. 20 mM aminoallyl dUTP (AA-dUTP): Dissolve 1 mg of AA-dUTP in 100 µL 0.2 M bicarbonate buffer.
3. 0.2 M Bicarbonate buffer: Dissolve 1.68 g of sodium bicarbonate ($NaHCO_3$) in 100 mL of distilled water. Sterilize by autoclaving and store at RT.
4. 2 M Glycine: Dissolve 75.07 g glycine in 0.5 L of distilled water. Sterilize by filtration through a 0.22 µm syringe filter and store at RT.
5. Cy3 Mono NHS Ester (Amersham): Dissolve 1 mg Cy3 Mono NHS Ester in 66 µL DMSO. Store at −20 °C.
6. Dig Mono NHS Ester (Invitrogen): Dissolve 5 mg Dig Mono NHS Ester in 213 µL DMSO. Store at −20 °C.
7. 1 M Tris–HCl (pH 7.75): Dissolve 12.11 g of Trizma Base in 50 mL distilled water. Adjust pH with HCl and bring the volume to 100 mL with distilled water. Sterilize by autoclaving and store at 4 °C.
8. 10 × NT Buffer: 0.5 M Tris–HCl (pH 7.5), 50 mM $MgCl_2$, 0.05% BSA (albumin bovine serum), ddH_2O. Keep 200 µl aliquots at −20 °C.

9. dNTPs mix: 0.5 mM dATP, 0.5 mM dCTP, 0.5 mM dGTP, and 0.1 mM dTTP. Keep 100 µl aliquots of mixed dNTPs at −20 °C.
10. 0.1 M mercaptoethanol: Add 50 µl mercaptoethanol to 7.2 mL SDW. Keep 500 µl aliquots at −20 °C.
11. DNA polymerase I (*E. coli*, Fermentas), 10 U/µl): Dilute 1:1 with DNA polymerase I storage buffer. Keep in −20 °C.
12. DNase I: Prepare stock 1 mg/mL by dissolving DNase I in 0.15 M NaCl in 50% glycerol. Prepare DNase I ready-to-use dilution (1:250) by mixing 2 µl DNase I (1 mg/mL) with 500 µl 0.15 M NaCl in 50% glycerol. Keep 25 µl aliquots at −20 °C.
13. 50% glycerol.
14. 0.15 M NaCl in 50% glycerol.
15. 1 M $MgCl_2$: Dissolve 10.16 g of $MgCl_2 \times 6H_2O$ in 50 mL of sterile ddH_2O.

2.3.3 Purification of Labeled Probes

1. 3 M sodium acetate, pH 5.2 (*see* **Note 4**).

   Dissolve 40.8 g sodium acetate·$3H_2O$ in 80 mL sterile ddH_2O. Adjust pH to 5.2 with glacial acetic acid and add ddH_2O to 100 mL. Sterilize by filtration through a 0.22 µm syringe filter. Keep at 4 °C or at RT.
2. Ethanol, ice-cold, 70 and 100% (*see* **Note 5**).
3. TE buffer: 10 mM Tris–HCl, pH 8, 1 mM EDTA, pH 8. Mix 10 mL of 1 M Tris–HCl and 2 mL of 0.5 M EDTA and add distilled water up to 1 L. This solution can be autoclaved or filter-sterilized through a 0.22 µm syringe filter and stored at 4 °C.
4. 0.5 M EDTA: Add 37.22 g of EDTA to 80 mL of distilled water. While stirring vigorously on a magnetic stirrer, add NaOH pellet or 10 N NaOH to adjust pH to 8.0. Add distilled H_2O to 100 mL. Sterilize by autoclaving and store at 4 °C.
5. 1 M Tris–HCl (pH 8.0): Dissolve 12.11 g of Trizma base in 80 mL of distilled H_2O. Add HCl to adjust pH to 8.0. Add distilled H_2O to 100 mL. Sterilize by autoclaving and store at 4 °C.
6. Salmon sperm blocking DNA.

2.4 Fluorescence In Situ Hybridization

1. 100 µg/mL RNase A in 2 × SSC: Prepare by diluting stock RNase A (10 mg/mL) in 2 × SSC in a proportion of 1:100.
2. 20 × SSC: 3 M NaCl, 0.3 M trisodium citrate dihydrate ($C_6H_5Na_3O_7{\cdot}2H_2O$). Dissolve 175.3 g NaCl and 88.3 g trisodium citrate dehydrate in 800 mL distilled water. Adjust pH

to 7.0 with 5 M HCl. Bring the volume to 1 L with distilled water and sterilize by autoclaving. Store at RT.

3. 2 × SSC: Prepare by diluting 20 × SSC with SDW. Store at RT.
4. 0.1 mg/mL pepsin solution in 0.01 M HCl (optional).
5. 10 × PBS: Prepare solution A using 0.1 M Na_2HPO_4 and 1.4 M NaCl and solution B using 0.1 M NaH_2PO_4 and 1.4 M NaCl. Dissolve 8.01 g Na_2HPO and 36.82 g NaCl in 450 mL distilled water to make solution A. Dissolve 1.56 g NaH_2PO_4 and 8.18 g NaCl in100 mL distilled water to make solution B. Add solution B to solution A until reaching pH 7.0. Sterilize by autoclaving. For use, dilute at a ratio of 1:10. Store at RT.
6. 1 × PBS: Prepare by dilution 10 × PBS with SDW. Store at RT.
7. 1% formaldehyde in 1 × PBS: Prepare by mixing 6 mL 37% acid-free formaldehyde, 20 mL 10 × PBS, and 174 mL SDW. This solution is unstable and should be prepared fresh just before use. Formaldehyde is toxic and should be handled in a fume hood.
8. 100% formamide, deionized: Mix 200 mL 100% formamide and 10 g Amberlite IRN-150 L monobed mixed resin. Stir vigorously for 1–2 h and filter. Store at −20 °C. Formamide is toxic and should be handled in a fume hood.
9. 50% dextran sulfate: Dissolve 5 g dextran sulfate in 10 mL SDW at 65 °C. Sterilize the solution by filtration through a 0.22 μm syringe filter (optional). Prepare 1 mL aliquots and store at −20 °C. (*see* **Note 6**)
10. 10% SDS: Dissolve 10 g SDS in 80 mL of sterile ddH_2O and then bring the volume to 100 mL. Store at RT.
11. Ethanol (70, 90 and 100%): Store at RT. Ethanol for FISH can be reused several times.
12. DAPI (100 μg/mL): Dissolve 1 mg DAPI (4′,6′-diamidino-2-phenylindole) in 10 mL ddH_2O. Aliquot to 1 mL and store at −20 °C.
13. Vectashield antifade mounting medium (Vector Laboratories) with 2.5 μg/mL DAPI: Mix 2.5 μl DAPI solution (100 μg/mL) and 97.5 μl Vectashield. Store the solution at 4 °C.

2.5 Immunodetection of Probes Labeled with Digoxigenin-dUTP

1. 4 × SSC + 0.2% Tween 20: Add 200 mL 20 × SSC and 2 mL Tween 20 to 800 mL SDW. Store at RT.
2. Blocking buffer (5% nonfat dry milk in 4 × SSC): Add 5 g nonfat dry milk and 20 mL 20 × SSC to 80 mL SDW. Mix the solution well and divide into 2 mL aliquots. Store at −20 °C.

3. FITC-conjugated anti-digoxigenin antibody: Dissolve whole lyophilisate in 1 mL ddH_2O. Prepare 30 µl aliquots and store in the dark at −20 °C. Before use, dilute in blocking buffer in a proportion of 1:11.

2.6 Equipment

1. Coplin jars.
2. Coverslips.
3. Dissection needles.
4. Dry ice or liquid nitrogen allowing rapid freezing of cytological preparations to at least −70 °C.
5. Eppendorf tubes.
6. Filter paper.
7. Fine forceps.
8. Glass Pasteur pipettes.
9. Image acquisition and processing system: epifluorescence microscope with a set of filters enabling excitation of DAPI, FITC, and tetramethyl-rhodamine fluorophores, high sensitivity CCD monochrome or color camera and a computer with software controlling image acquisition and processing.
10. Incubator at 37 °C.
11. In situ thermal cycler (e.g., Hybaid OmniSlide System, Thermo Scientific).
12. Laboratory microscope with ×10, ×20, and ×40 phase contrast objectives.
13. Laboratory stereoscopic dissecting microscope with incident light.
14. Microcentrifuge.
15. Microscope slides.
16. PCR thermal cycler.
17. Petri dishes (60 and 90 mm).
18. pH-meter.
19. Plastic coverslips, thermostable (*see* **Note 7**).
20. Plastic tubes, volume 25 mL.
21. Scalpels.
22. Thermal mixer (optional).
23. Watch glasses.
24. Water bath.

3 Methods

3.1 *Brachypodium* Mitotic/Meiotic Material Preparation

1. Germinate the seeds for 3–5 days in the dark in Petri dishes on filter paper moistened with distilled water.
2. When the seedlings have roots 1.5–2.0 cm long, use ca. half of them for preparing mitotic material. Immerse the seedlings in ice-cold water in 50 ml plastic tubes and incubate for 24 h at 4 °C in order to accumulate metaphases. The other ca. half of the seedlings will be the source of meiotic material.
3. After the treatment with ice-cold water, fix the seedlings in Carnoy's fixative at RT for several hours and then store at −20 °C until required.
4. Sow the rest of the seedlings in pots filled with soil mixed with vermiculite at a ratio of 3:1. Grow the plants in a greenhouse under 16 h days at 20 ± 1 °C and illuminate by lamps emitting white light with an intensity of 10,000 lx.
5. In order to induce synchronized flowering, subject 4-week-old plants to vernalization for 3 weeks at 4 °C. After the period of vernalization, return the plants to the greenhouse. They should flower within 3–5 weeks.
6. Collect immature spikes from plants, fix them in Carnoy's fixative, and store at −20 °C.

3.2 Mitotic and Meiotic Chromosome Preparations

3.2.1 Mitotic Chromosome Preparations

1. Rinse fixed root meristems for 3 × 5 min in a small Petri dish containing 0.01 M citrate buffer.
2. Replace the 0.01 M citrate buffer with the enzyme mixture for digesting roots, cover the Petri dish with a watch glass, and digest the root meristems for 2 h at 37 °C.
3. Transfer one root meristem to a small watch glass containing 2–3 mL of 45% acetic acid.
4. Using fine needles remove the root cap and extrude the meristem under a dissecting microscope. Transfer the extruded meristem into a drop of 45% acetic acid in the center of a clean slide, apply a coverslip, and gently squash the preparation.
5. Check the quality of the preparation under a phase contrast microscope. A high-quality preparation contains many (at least 10–15) well-spread metaphase plates with condensed, non-overlapping chromosomes.
6. Freeze the high-quality slides for about 30 min by placing them in a container with dry ice or plunge the slides, one by one, into liquid nitrogen holding them by tweezers for 5–10 s. Remove the coverslips with a razor blade and air-dry the slides.

3.2.2 Meiotic Chromosome Preparations

1. Place fixed immature spikes in a Petri dish lined with filter paper moistened with 0.01 M citrate buffer.
2. Isolate anthers using fine needles or scalpels, transfer them to a small Petri dish containing 2 mL of 0.01 M citrate buffer, and rinse for 15 min (*see* **Note 8**). Due to the relatively low gradient of the meiotic phases in the spikes of *Brachypodium* species, the preparations are made without initial meiotic phase screening. Each spikelet of *B. distachyon* contains two anthers; all other species have three anthers per spikelet.
3. Replace the 0.01 M citrate buffer with the enzyme mixture, cover the Petri dish with a watch glass, and digest the anthers for 2 h at 37 °C.
4. Carefully transfer digested anthers to one drop of 45% acetic acid on a clean slide, break anthers into small pieces by needles, apply a coverslip, and squash. Use three to four anthers for a single slide.
5. Freeze the preparations on dry ice or in the liquid nitrogen, flick off the coverslips with a razor blade, and air-dry the slides.

3.3 BAC DNA Isolation

BD_ABa and BD_CBa *B. distachyon* genomic DNA libraries [3, 4] can be the source of the BAC clones used for the chromosome barcoding and chromosome painting of *Brachypodium* species. BAC clones linked to centromeric and pericentromeric regions should be excluded from the experiment since they can yield non-specific cross-hybridization signals on chromosomes other than the targeted one. For similar reasons, using clones that contain more than 30% of repetitive sequences should be avoided. The list of the BAC clones suitable for cytogenetic mapping of chromosomes of *Brachypodium* species can be requested from the authors.

1. Prepare sterile Petri dishes containing ~25–50 mL of LB Agar supplied with 12.5 μg/mL chloramphenicol.
2. Inoculate bacteria using streak-plate procedure to obtain pure single colonies. Use a flame-sterilized metal loop or sterile wooden toothpicks to transfer bacterial samples from a micro-titer plate and spread it over the surface of the solidified LB Agar medium.
3. Place closed Petri dishes upside down in the incubator set to 37 °C and incubate for 16 h. Prolong the incubation time, if necessary, until the single colonies reach the size of 1–2 mm.
4. Transfer the selected colonies from the Petri dishes into sterile tubes containing 10 mL of LB Broth with 12.5 μg/mL of chloramphenicol and incubate in a shaking incubator at 37 °C, 250 rpm overnight (~16 h).
5. For CB or CCB, isolate the DNA of each clone separately. For CP or CCP, divide selected BACs into pools of eight to ten

clones each and mix equal volumes of the bacterial cultures (*see* **Note 9**). Perform BAC DNA isolation as described by Kotchoni et al. [11] with minor modifications.

6. Harvest the cells by centrifugation of the bacterial culture in 5 mL Eppendorf tubes at 4000 × *g* for 5 min at RT (*see* **Note 10**).
7. Resuspend a bacterial pellet in 300 μl of solution I (*see* **Note 3**) containing 100 μg/mL of RNase A, vortex. Incubate for 10 min at RT.
8. Add 300 μl of freshly prepared solution II (*see* **Note 3**) and mix well by gently inverting it four to six times. Incubate for 5 min at RT.
9. Add 300 μl of solution III (*see* **Note 3**) and mix the content of the tubes very gently by pipetting. Incubate the tubes on ice for 10 min without shaking.
10. Centrifuge the tubes at 10,000 × *g* for 10 min at RT and then carefully transfer the supernatant into new Eppendorf tubes.
11. Add 650 μl of isopropanol to the supernatant. Incubate the tubes for 2 min at RT and centrifuge at 14000 × *g* for 20 min.
12. Wash the pellet twice in 500 μl of 70% ethanol at RT and centrifuge at 14000 × *g* for 5 min. Air-dry after removing the ethanol.
13. Dissolve a dried pellet in 30 μl of buffer EB. Incubate overnight at 4 °C.
14. Store the BAC-DNA solutions at −20 °C.

3.4 Probe Labeling

3.4.1 Probe Labeling Using Commercially Available Nucleotides

1. Centrifuge briefly dATP, dCTP, dGTP, dTTP, and the labeled nucleotides.
2. Combine the reagents in an Eppendorf tube placed on ice according to Table 1.
3. Carry out the nick-translation reaction in the PCR thermal cycler under the following conditions: 15 °C × 95 min + 65 °C × 10 min + 4 °C × ∞.

3.4.2 Probe Labeling Using Custom-Made Labeled Nucleotides

1. Prepare custom-made labeled nucleotides according to Henegariu et al. [12] in a two-step procedure. Use this procedure to prepare digoxigenin-dUTP and Cy3-dUTP. The first step for the generation of digoxigenin-dUTP/Cy3-dUTP is presented in Table 2.

   The formula for calculating the volume of the solvent required to dissolve the dyes, haptens, or allylamine in DMSO is as follows:

Table 1
BAC-DNA labeling mixture with commercially available labeled nucleotides

Components	Quantity (μl)
0.4 mM dATP	2.5
0.4 mM dCTP	2.5
0.4 mM dGTP	2.5
0.4 mM dTTP	1.67
tetramethylrhodamine-5-dUTP or digoxigenin-11-dUTP	0.83
BAC DNA 100–500 ng/μl	up to 6
Nick Translation Mix (Roche)	4
ddH_2O	Water up to final volume of 20 μl

Table 2
The first step for the generation of digoxigenin-dUTP/Cy3-dUTP

	Quantity (μl)	
Components	Digoxigenin-dUTP	Cy3-dUTP
20 mM AA-dUTP	10	10
0.2 M bicarbonate buffer	10	10
DMSO	10	0
40 mM DIG mono NHS ester or 20 mM Cy3 mono NHS ester	10	10
H_2O	15	10
	$\sum = 55$ μl	$\sum = 40$ μl

$$\text{Solvent } (\mu l) = \frac{\text{Reagent (mg)}}{\text{MW} \times \text{desired molarity (mM)}} \times 1,000,000$$

2. Incubate the samples for 4 h at RT. Add the components listed in Table 3 to the reagents from the first step.
3. Label the BAC DNA using nick translation. The composition of the nick-translation reaction mixture is shown in Table 4.

Table 3
The second step for the generation of digoxigenin-dUTP/Cy3-dUTP

Components	Quantity (μl)	
	Digoxigenin-dUTP	Cy3-dUTP
2 M glycine (pH 8)	2	2
1 M Tris–HCl (pH 7.75)	4	4
H_2O	139	154
	$\sum = 200$ μl	$\sum = 200$ μl

Table 4
The reaction mixture for BAC DNA labeling using custom-made nucleotides

Components	Quantity (μl)
NT Buffer	2.5
2 mM dNTPs mix	2.5
0.1 M mercaptoethanol	2.5
Custom-labeled digoxigenin-dUTP/Cy3-dUTP	2.0
DNA polymerase 10 U/μl	2
DNase I	1.5
BAC DNA	0.5
ddH_2O	Water up to final volume of 25 μl

4. Carry out the nick-translation reaction in the PCR thermal cycler under the following conditions: 15 °C × 110 min + 65 °C × 10 min + 4 °C × ∞

3.4.3 Purification of Labeled Probes

Precipitate the labeled probes in ethanol in order to remove the residues of the labeling mixture and to condense the products (*see* **Notes 11** and **12**).

1. Add 0.1× probe volume of 3 M sodium acetate pH 5.2 and 2.5× probe volume of ice-cold 100% ethanol.
2. Incubate the probe at −80 °C for 30 min or −20 °C overnight.
3. Centrifuge the probe at 14000 × *g* for 30 min at 4 °C and pipette off the supernatant.
4. Wash the pellet in 2.5× probe volume of ice-cold 70% ethanol. Centrifuge at 14000 × *g* for 10 min at 4 °C.

5. Pipette off the supernatant and dry the DNA pellet in the incubator at 37 °C for 10–12 min.
6. For CB and CCB resuspend the BAC DNA pellet in TE buffer for 30 min at RT and then overnight at 4 °C. Store the labeled probes in −20 °C until use.
7. For CP and CCP resuspend the BAC DNA pellet in a hybridization mixture for 2–3 h at 37 °C and use immediately for FISH (*see* **Note 13**).

3.5 Fluorescence In Situ Hybridization

3.5.1 Slide Pretreatment

1. Add 200 µl of 100 µg/mL RNase A in 2 × SSC to each slide and cover with a plastic coverslip. Incubate the slides for 1 h at 37 °C in a humid chamber.
2. Wash the slides three times for 5 min each in 2 × SSC in at RT (*see* **Note 14**).
3. (Optional) In order to remove excess cytoplasm from the meiotic chromosome preparations, after treatment with RNase, incubate the slides in a solution of pepsin diluted in 0.01 M HCl (0.1 mg/mL) for 10 min at 37 °C. Wash the slides three times for 5 min each in 2 × SSC at RT.
4. Place the slides for 10 min in freshly prepared 1% formaldehyde in 1 × PBS at RT.
5. Wash three times for 5 min each in 2 × SSC at RT.
6. Dehydrate the slides in successive washes of 70, 90, and 100% ethanol (3 min each wash) and air-dry at RT.

3.5.2 Denaturation, Hybridization, and Stringent Washing

1. Prepare the hybridization mixture with labeled BAC DNA probes in the Eppendorf tubes placed on ice. For CB and CCB, the BAC DNA of individual clones after precipitation is dissolved in TE buffer and added to the hybridization mixture described in Table 5. When using the pools of BAC clones for CP and CCP, it is recommended to dissolve the precipitated probe directly in the hybridization mixture (Table 6) in order to ensure high concentration of BAC DNA and the efficiency of FISH. The kinetic power for both the hybridization mixture and stringent washing equals 79% for CB/CP and 59% for CCB/CCP [13].
2. Predenature the probes in the hybridization mixture for 10 min at 75 °C and then immediately plunge the samples into ice for 10 min in order to stabilize the probe.
3. Add 38 µl of the hybridization mixture to each slide and cover with a plastic coverslip.
4. Denature the slides in an in situ thermal cycler (Hybaid Omni-Slide, Thermo Scientific) for 4 min and 30 s at 70 °C (*see* **Note 15**).

Table 5
Hybridization mixture used for CB and CCB

Component	Quantity (μl)	
	CB	CCB
100% formamide	20	16
50% dextran sulfate (*see* **Note 6**)	8	12
20 × SSC	4	4
10% SDS	2	2
BAC DNA probe 1 (dissolved in TE)	Up to 3 (100–250 ng/slide)	Up to 3 (100–250 ng/slide)
BAC DNA probe 2 (dissolved in TE)	Up to 3 (100–250 ng/slide)	Up to 3 (100–250 ng/slide)
ddH_2O	Water up to final volume of 40 μl	Water up to final volume of 40 μl
	∑ = 40 μl	∑ = 40 μl

Table 6
Hybridization mixture used for CP and CCP

Component	Quantity (μl)	
	CP	CCP
100% formamide	20	16
50% dextran sulfate	8	12
20× SSC	4	4
H_2O	8	8
	∑ = 40 μl	∑ = 40 μl

5. Transfer the slides to a moist chamber and incubate at 37 °C. The incubation time is ~24 h in the case of CB/CP in *B. distachyon* and ~48 h in the case of CCB/CCP in all other species.
6. In order to perform stringent washing, follow the steps listed in Table 7. If your probes were labeled with rhodamine-dUTP only, proceed to **step 7** of the Subheading 3.5.2. If the probes were labeled with digoxigenin-dUTP, proceed directly to the immunodetection of digoxigenated DNA probes (Subheading 3.5.3).

Table 7
Different parameters of stringent washing used in the experiments

Steps	Stringency parameters	CB/CP	CCB/CCP
I	2 × SSC	5 min × 3, 42 °C	5 min × 3, 37 °C
II	20% (vol/vol) formamide in 0.1 × SSC for CB/CP 10% (vol/vol) formamide in 2 × SSC for CCB/CCP	10 min, 42 °C	10 min, 37 °C
III	2 × SSC	5 min × 2, 42 °C	5 min × 2, 37 °C
IV	2 × SSC	5 min × 2, RT	5 min × 2, RT

7. Dehydrate the slides in successive washes of 70, 90, and 100% ethanol (3 min each wash) and air-dry at RT.
8. Counterstain the slides with 2.5 μg/mL DAPI in Vectashield and store at 4 °C in the dark.

3.5.3 Immunodetection of Probes Labeled with Digoxigenin-dUTP

1. Wash the slides in 4 × SSC + 0.2% Tween 20 for 5 min at RT.
2. Apply 200 μl of a blocking buffer (5% skimmed milk solution in 4 × SSC) to each slide. Cover the slides with plastic coverslips and then transfer them to a moist chamber for 30 min at RT.
3. Gently remove the coverslips using forceps. Apply 40 μl of FITC-conjugated anti-digoxigenin diluted in a blocking buffer (1:11) to each slide. Apply new plastic coverslips, place the slides in a moist chamber, and incubate for 1–2 h at 37 °C.
4. Wash the slides three times in 4 × SSC with 0.2% Tween 20 (8 min each wash) at 37 °C.
5. Dehydrate the slides in a 70, 90, and 100% ethanol series (3 min in each) and air-dry at RT.
6. Counterstain the slides with 2.5 μg/mL DAPI in Vectashield and store at 4 °C in the dark.

3.5.4 Microscopic Analysis and Documentation of Results

Analyze the slides using an epifluorescent microscope equipped with monochromatic of color digital charge-coupled device (CCD) camera and software controlling image acquisition. Use appropriate filters for the following fluorochromes: DAPI (maximum absorption 350 nm); fluorescein isothiocyanate (maximum absorption 494 nm); tetramethyl-rhodamine (maximum absorption 555 nm). Process the captured images and digitally superimpose them using software of choice, e.g., Photoshop (Adobe) or ImageJ (NIH).

4 Notes

1. Alternatively, ethanol could be used instead of methanol.
2. The LB media should be cooled down to ~50–60 °C before adding chloramphenicol since an overly high temperature will cause thermal degradation of the antibiotic. However, chloramphenicol has to be added to LB Agar before the medium starts to solidify. It is safe to add chloramphenicol if the bottom of the flask containing LB Agar is not too hot to touch.
3. Instead of preparing solutions I, II, and II, commercially available buffers for plasmid DNA isolation (P1, P2, P3 or N3, e.g., QIAGEN) can be used.
4. For purification of DNA probes labeled with the use of custom-made nucleotides, use 3 M sodium acetate, pH 5.6.
5. For purification of DNA probes labeled with the use of custom-made nucleotides, use 80 and 100% ethanol.
6. Dextran sulfate is very viscous. It is recommended to warm it up to ~65 °C in a thermal mixer or water bath before pipetting.
7. Plastic autoclavable bags cut into 24 mm × 24 mm pieces are conveniently used.
8. In order to transfer the anthers, they should be placed on the tip of a needle or scalpel. If transferred by a micropipette, they can become stuck to the inside of the plastic tip.
9. Even though the clones for CP and CCP can be isolated and then labeled separately, pooling them during the DNA isolation stage greatly reduces the time and cost of the procedure. The number of clones constituting a pool can be adjusted to the researcher's needs.
10. 2.2 mL Eppendorf tubes can be used instead of 5 mL Eppendorf tubes. In such a case, centrifuge 2 mL of the bacterial culture at 4000 × *g* for 5 min at RT and discard the supernatant. Repeat the step twice until harvesting the cells from the total culture volume of 5–6 mL.
11. For CP and CCP, labeled samples could be combined at this point in order to prepare painting BAC pools. Mix equal volumes of the samples before adding 3 M sodium acetate and 100% ice-cold ethanol. Pooled probes comprising up to 150 BAC clones have been proved to yield satisfying hybridization signals.
12. For purifying the probes labeled with custom-made nucleotides, add 36 μl of 3 M sodium acetate pH 5.6, 1 mL of ice-cold 100% ethanol, and 2 μl of salmon sperm blocking DNA to the sample. Incubate the probes at −80 °C for 30 min and centrifuge at 14000 × *g* for 30 min at 4 °C. Add one mL of ice-cold

80% ethanol and then centrifuge the tubes at 14000 × *g* for 10 min at 4 °C. Pipette off the supernatant and air-dry the DNA at RT. Then proceed to **step 6** in the Subheading 3.4.3.

13. Resuspending the BAC DNA in the hybridization mixture can be sped up by placing the samples in a thermal mixer at 37 °C and rotating at 300–350 rpm.
14. Use glass Coplin jars to wash the slides during pretreatment, stringent washing, and immunodetection of digoxigenin-labeled probes. For handling higher numbers of slides (up to 25), use plastic dishes for cytological staining equipped with slide racks.
15. Instead of an in situ thermal cycler, an incubator set to 70 °C can be used.

Acknowledgment

This work was supported by the Polish National Science Centre [grant no. DEC-2012/04/A/NZ3/00572]. DI-H acknowledges the 'Grant for Young Scientists' awarded by the University of Silesia Faculty of Biology and Environmental Protection.

References

1. Garvin D, Gu Y, Hasterok R, Hazen S, Jenkins G, Mockler T, Mur L, Vogel J (2008) Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. Crop science. Plant Genome 48:69–84
2. Vogel JP, Bragg J (2009) *Brachypodium distachyon*, a new model for the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) Plant genetics and genomics: crops and models, vol vol 7. Springer Verlag, New York, pp 427–449
3. Febrer M, Goicoechea JL, Wright J, McKenzie N, Song X, Lin J, Collura K, Wissotski M, Yu Y, Ammiraju JS, Wolny E, Idziak D, Betekhtin A, Kudrna D, Hasterok R, Wing RA, Bevan MW (2010) An integrated physical, genetic and cytogenetic map of *Brachypodium distachyon*, a model system for grass research. PLoS One 5(10):e13461. doi:10.1371/journal.pone.0013461
4. International_Brachypodium_Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463 (7282):763–768. doi:10.1038/nature08747
5. International_Brachypodium_Initiative (2014) Update on the genomics and basic biology of Brachypodium. Trends Plant Sci 19 (7):414–418. doi:10.1016/j.tplants.2014.05.002
6. Idziak D, Betekhtin A, Wolny E, Lesniewska K, Wright J, Febrer M, Bevan MW, Jenkins G, Hasterok R (2011) Painting the chromosomes of *Brachypodium*: current status and future prospects. Chromosoma 120(5):469–479. doi:10.1007/s00412-011-0326-9
7. Betekhtin A, Jenkins G, Hasterok R (2014) Reconstructing the evolution of *Brachypodium* genomes using comparative chromosome painting. PLoS One 9(12):e115108. doi:10.1371/journal.pone.0115108
8. Hasterok R, Marasek A, Donnison IS, Armstead I, Thomas A, King IP, Wolny E, Idziak D, Draper J, Jenkins G (2006) Alignment of the genomes of *Brachypodium distachyon* and temperate cereals and grasses using bacterial artificial chromosome landing with fluorescence *in situ* hybridization. Genetics 173 (1):349–362. doi:10.1534/genetics.105.049726
9. Wolny E, Lesniewska K, Hasterok R, Langdon T (2011) Compact genomes and complex evolution in the genus *Brachypodium*. Chromosoma 120(2):199–212. doi:10.1007/s00412-010-0303-8

10. Idziak D, Hazuka I, Poliwczak B, Wiszynska A, Wolny E, Hasterok R (2014) Insight into the karyotype evolution of *Brachypodium* species using comparative chromosome barcoding. PLoS One 9(3):e93503. doi:10.1371/journal.pone.0093503
11. Kotchoni SO, Gachomo EW, Betiku E, Shonukan OO (2003) A home made kit for plasmid DNA mini-preparation. Afr J Biotechnol 2:88–90
12. Henegariu O, Bray-Ward P, Ward DC (2000) Custom fluorescent-nucleotide synthesis as an alternative method for nucleic acid labeling. Nat Biotechnol 18(3):345–348. doi:10.1038/73815
13. Schwarzacher T, Heslop-Harrison JS (2000) Practical in situ hybridization. BIOS Scientific, Milton Park, England

Chapter 2

Transcriptional and Posttranscriptional Regulation of Drought Stress Treatments in *Brachypodium* Leaves

Edoardo Bertolini, Mario Enrico Pè, and Erica Mica

Abstract

Plant sensing drought stress conditions activate complex molecular networks leading to a rapid reprogramming of plant physiology and metabolism, in order to survive in suboptimal conditions.

Here, we describe a standardized in vivo soil drought assay to investigate the effects of drought stress on leaf growth. Since it is now clear that stress responses can be specific to developmental stages and cell types, we describe a procedure to dissect the leaf in three distinct areas in order to study transcriptional and posttranscriptional gene regulation on both organ and cellular levels. Noncoding RNAs, both small RNAs and long noncoding RNAs, are emerging to be deeply involved in abiotic stress responses, acting as molecular switches, interconnecting different response pathways. Here, we illustrate the methodology that has been used to identify miRNAs involved in drought response and to analyze the modulation of expression of their putative targets, in order to gain a complete picture of transcriptional and posttranscriptional regulation driven by noncoding RNAs.

Key words miRNAs, Drought, *Brachypodium*, Leaf development, Next-generation sequencing

1 Introduction

Temperate cereals are the most important crops worldwide but lack of simple genetics and genomics tools because of their large genomes, breeding programs are time consuming with several experimental limitation. In recent years, the rapid advance of molecular biology, transgenesis and functional genomics applied to the model species *Brachypodium distachyon* (Bd) have facilitated significant progress in identifying crucial aspects of plant biology that may boost applied and translational research projects [1].

Relevant for the scope of this method is the fact that Bd, originated in Iraq [2], is extremely drought-tolerant and possesses specific adaptations mechanisms which may be transferred to related grass species, such as barley and wheat.

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_2,

It is now accepted that plants actively reduce their growth when they encounter stress, in order to redistribute their resources and appropriately respond to the abiotic stress by deploying various drought tolerance mechanisms [3]. The underlying molecular mechanisms and pathways that coordinate growth control and drought tolerance are still poorly understood especially in monocotyledonous plants. The identification and study of noncoding RNAs (ncRNAs), including microRNAs and small RNAs, as well as the new emerging class of long noncoding RNAs, has added a new layer of complexity to the pathways that regulate plant development and drought stress. These molecules, which function as negative and positive regulators of gene expression at both transcriptional and posttranscriptional level, are now known to have greatly expanded their role in a variety of developmental processes, and recent studies reported that several ncRNAs are associated with abiotic stress responses [4].

Here, we describe a method to set up a high reproducible drought protocol that has been optimized to subject Bd plants to an in vivo, non-lethal, growth-limiting drought stress which can be easily adapted to other monocot species. The second part of the chapter shows the procedure to dissect the leaf in three distinct areas (three cell types of the third leaf): proliferating, expanding, and mature cells, grown under optimal conditions, intermediate mild drought, and severe drought stress, and to extract RNA and the small RNA fraction, in order to study transcriptional and post-transcriptional gene regulation at both organ and cellular levels.

2 Materials

2.1 Soil Drought Experiment

1. *Brachypodium* seeds (*see* **Note 1**).
2. Round petri dishes (145/20 mm).
3. Distilled and tap water.
4. Peat dish pellets.
5. Potting compost (*see* **Note 2**).
6. Small plastic pots (5.5 cm diameter, 5 cm high) (*see* **Note 3**).
7. ARATRAY (http://www.arasystem.com/).
8. Surgical tape.
9. Aluminum foil.
10. P1000 micropipette.
11. Digital scientific scale.
12. Razor blades.
13. Forceps.
14. RNA*later*® stabilization solution (Thermo Fisher Scientific).

15. Liquid nitrogen.
16. Data logger for recording temperature and humidity.
17. Growth room (*see* **Note 4**).

2.2 Total RNA and Small RNA Extraction

1. Mortar and pestle.
2. Agarose gel (*see* **Note 5**).
3. Nanodrop Spectrometer (Thermo Fisher Scientific).
4. Agilent 2100 Bioanalyzer (Agilent Technology).
5. mirPremier™ microRNA Isolation Kit (Sigma-Aldrich).
6. Plant/Fungi Total RNA Purification Kit (Norgen Biotek).
7. TruSeq Small RNA Sample Preparation Kits (Illumina).
8. TruSeq stranded mRNA kit (Illumina) (*see* **Note 6**).

2.3 Sequencing Approach

1. Multiplexed 50 bp Single-read (SR) sequencing.
2. Multiplexed 100 bp Single-read (SR) sequencing.

2.4 Data Analysis

1. Reference Genome.
2. ShortStack (*see* **Note 7**).
3. TargetFinder (*see* **Note 8**).
4. pstMimic (*see* **Note 9**).

3 Methods

This chapter aims at describing the detailed protocol to set up a high reproducible soil drought assay that is non-lethal, i.e., *Brachypodium* is able to complete its life cycle through all the phenological stages (Fig. 1). In this methods, plants of Bd inbred line 21 are subjected to two different levels of drought stress expressed in terms of relative soil water content (RSWC) (*see* **Note 10**), which reflects the water holding capacity of the soil. Optimal condition, intermediate mild stress and severe drought stress are set respectively to 83, 33, and 20% RSWC (Fig. 2). These drought stress levels are particularly appropriate for Bd, which is a drought tolerant grass originated in Iraq. Even though this protocol have been developed for Bd, in our laboratory this method has been also applied with great success to some *Triticum species* (unpublished data). In this method, we show how to collect the third growing leaf and its three developmental zones that will be used for profiling small RNAs and mRNAs.

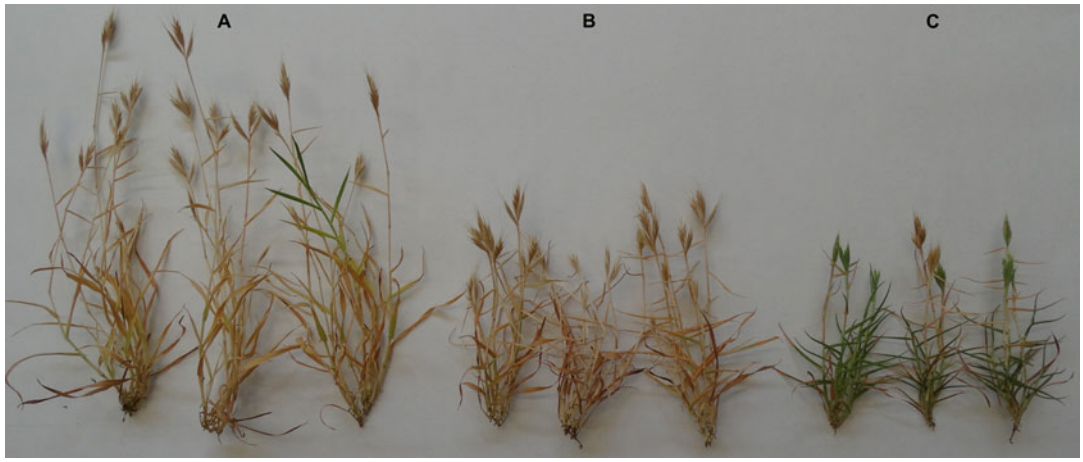

Fig. 1 Brachypodium Bd21 plants at maturity, 90 DAG (days after germination), grown in control condition (A), moderate drought condition (B), and severe drought condition (C)

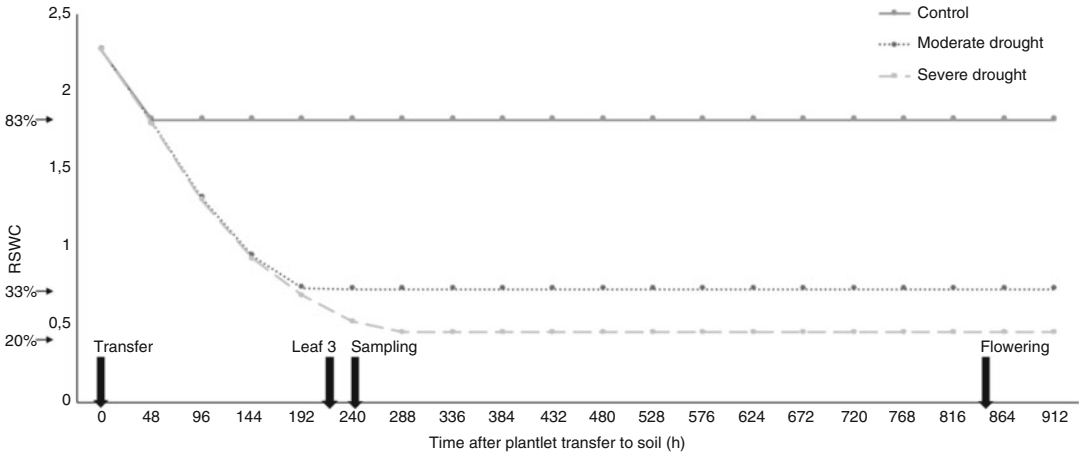

Fig. 2 Progression of soil drought experiment showing the reproducibility of the assay. *X*-axis represents the time (in hours) from plantlets transfer, whereas *Y*-axis represents grams of water per grams of soil; arrows indicate the relative water soil content (RSWC) of the three stress conditions

3.1 Drought Protocol

3.1.1 Pots Calibration

1. Air-dry the compost to remove the residual humidity for ~15 days.
2. In each pot add 12 g of air-dried compost and calculate the grams of water to reach 100% RSWC.
3. One day before the transfer of plantlets (**step 8** in Subheading 3.1.2), calibrate the pots adding the grams of tap water to reach 100% RSWC (*see* **Note 11**). Calibrate the pots daily and adjust the weight immediately before plantlets transfer.

3.1.2 Standardized Growth Protocol

1. Transfer Bd21 seeds at 4 °C for at least 2 weeks before beginning the experiment.
2. Soak with distilled water two peat pellets until saturation.
3. Take out the soil from the peat pellet and spread it evenly on the petri dish.
4. Remove manually lemma and palea and place the seeds with the embryo in contact with the wet soil using a forcep. Carefully arrange the seeds equally distant from each other on the surface.
5. Seal the petri dish with surgical tape and incubate at 4 °C for 2 days in darkness. Use aluminum foil to completely cover the petri dish. This stratification step will improve the synchronous germination.
6. Transfer the plate to a growth chamber, with 16 h of light, at 24 °C and 55% relative humidity. Set the datalogger with these parameters and monitor the growth chamber condition twice a week.
7. After 3 days, when all seeds have germinated synchronously, remove the lid to let plantlets accustom to the ambient atmosphere in the growth chamber.
8. When the first leaf is about 3 cm in length, individual plantlets are carefully transferred to calibrated pots with 100% RSWC.
9. From this step on, control plants are dried down to 83% RSWC and this level will be kept for the whole experiment as control condition, while stressed plants are no longer watered, until they reach the desired RSWC.
10. Drought plants subjected to moderate stress are dried down to 33% RSWC.
11. Drought plants subjected to severe stress are dried down to 20% RSWC.
12. From this step on, monitor pot weight on a daily basis every 12 h and, to compensate for evapotranspirative losses, add water when necessary to readjust the pot weight to the target level (i.e., the desired RWSC). Take care not to deposit the water in the direct vicinity of a plant, but rather on the outer edges of the pots, and water always at the same time during the afternoon, to not interfere with the time of most active plant growth.
13. Three days later, the second leaf emerges in a highly reproductive manner, and another 3 days later the third leaf appears.
14. Nearly all plants should grow synchronously, and the third leaf of plants within the same experiment always appears within a 24h time window.

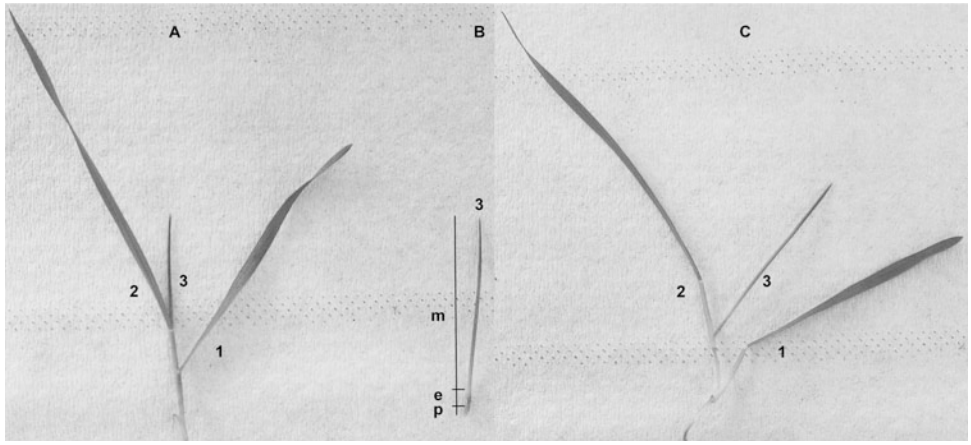

Fig. 3 Bd21 plant grown in severe drought condition and collected at 240 h from the plantlets transfer. (**A**) shows plant seedling at three leaves stage; (**B**) shows the third leaf collected and the three developmental leaf zones: *p* proliferation, *e* expansion, and *m* mature; (**C**) sampling of the third leaf from the leaf sheath to carefully collect proliferation, expansion and mature leaf zones. Numbers indicate the number of leaves

3.1.3 Third Leaf Sampling for Molecular Analysis

1. Harvest plants at a fixed time point in early afternoon, about 24 h after the emergence of the third leaf. The growing third leaf—between 1.5 and 2 cm in size (Fig. 3) at that point—is carefully removed from the leaf sheath of the older two leaves, without damaging the fragile meristem at its base.
2. Store immediately samples in tubes containing 5 ml of RNA-*later*® stabilization solution.
3. After an overnight incubation at 4 °C, dissect leaves into three distinct developmental zones with a sharp razor blade. Based on microscopic observations we defined the proliferation zone as the first 2 mm from the leaf base, the expansion zone as the next 4 mm, and the mature zone as the remaining distal part of the leaf (Fig. 3). To confirm these microscopic obsevation and the correct separation of the three developmental zones perform a RT-qPCR on several molecular markers specific for the proliferation and expansion zones [5].
4. Froze the leaf zones at −80 °C and store until RNA extraction.

3.2 Total RNA and Small RNA Extraction

1. Gently grind the three leaf zones in liquid nitrogen using pestle and mortar.
2. Extract the small RNA fraction using mirPremier™ microRNA Isolation Kit (Sigma-Aldrich) according to the manufacturer's protocol.
3. Extract total RNA using Plant/Fungi Total RNA Purification Kit (Norgen Biotek) according to the manufacturer's protocol.

4. Check quality and quantity of total RNA and small RNA with a NanoDrop spectrometer (Thermo Scientific) and by running a small aliquot on agarose gel (*see* **Note 5**).

3.3 RNA Sequencing

1. Generate small RNA libraries from each sample using TruSeq Small RNA Sample Preparation Kit (Illumina) with a multiplexing strategy.
2. Generate whole transcriptome libraries using TruSeq stranded mRNA kit (Illimina) (*see* **Note 6**).
3. Assess quality of the library using Agilent 2100 Bioanalyzer (Agilent Technologies)
4. Perform small RNA sequencing using a multiplexed approach generating for each sample 20 million 50 bp SR raw reads.
5. Perform whole transcriptome sequencing using a multiplexed approach with a 100 bp SR strategy. To detect all the expressed transcripts with the aim of characterizing also the low abundance transcript 60 million raw reads can be generated for each sample.

3.4 Data Analyses

Small RNA-seq and mRNA-seq data analysis start with a quality check of the reads and by removing the sequencing adapters (see Chapter 3).

1. Trimmed small RNA reads are processed using ShortStack program, developed by Michael J. Axtell [6], able to a make a comprehensive and informative annotation of known and novel MIRNA genes and phased small RNAs.
2. mRNA-seq data are processed using the bioinformatic approach described in the long noncoding RNAs protocol (see Chapter 3), able to detect the expression level of both coding and noncoding transcripts.
3. In order to highlight putative miRNA targets within the expressed transcripts, a target analysis can be performed by using TargetFinder program developed by Carrington laboratory [7]. This program will computationally predict small RNA binding sites on target transcripts providing a base-pairing diagram of the target and small RNA sequences.
4. To investigate the properties of endogenous long noncoding RNAs able to negatively regulate miRNA activities, the program pstMimic can be applied. This software, developed by Omicslab [8], is able to predict plant miRNA endogenous target mimics with sequence complementarity to the miRNA. Typically all the predicted target mimics show a bulge around 10–11 nt (the supposed miRNA cleavage site) on miRNA sequences.

4 Notes

1. To improve the germination do not use seeds older than 2 years and store the seeds at 4 °C.
2. The detailed potting compost composition used in this methods is: 50% black peat, 50% white peat, 20% organic substance, pH 5–6.5, dry material 25%, fertilizer NPK 12 + 14 + 24 1.5 Kg/m^3.
3. Small plastic cups (diameter 5.5 cm and height 5.5 cm) can be used instead of standard pots. Importantly they are closed at the bottom to prevent water loss. Moreover is better to use cup/pot that fit into the ARATRAY to maintain the plastic cup stable.
4. Growth room settings are: 16 h of light and 8 h of dark at 24 °C with 55% relative humidity.
5. Run an aliquot of total RNA and small RNA respectively on 1 and 3% w/v agarose gel.
6. Among the large number of transcripts transcribed in an eukaryotic cell, coding RNAs represent only 1–2% while the majority are noncoding RNAs (ncRNA). For this reason to dissect the complete transcriptome of both polyadenylated and non-polyadenylated RNAs, the Illumina Ribo-Zero rRNA kit (Plant Leaf) can be employed to remove cytoplasmic, chloroplastic and mitochondrial ribosomial RNAs. In this case perform a RiboZero treatment and proceed with the library preparation by skipping the PolyA selection step in the TruSeq stranded mRNA kit (Illimina). This variant has been applied successfully in several plant species (http://www.illumina.com/products/rrna-removal-kit-species-compatibility.html).
7. ShortStack can be obtained from Github (https://github.com/MikeAxtell/ShortStack/releases/) or can be executed using the web interface from the iPlant Discovery Environment platform (http://www.iplantcollaborative.org/ci/discovery-environment).
8. TargetFinder can be executed locally by downloading the script from Github (https://github.com/carringtonlab/TargetFinder). The code repository contains also a script version able to use multiple threads. Set the prediction score cutoff value equal to 3.
9. The pstMimic source code can be downloaded directly from the Omics lab homepage (http://omicslab.genetics.ac.cn/pstMimic/)

10. RSWC: the ratio of current soil water content to water content at field capacity, multiplied by 100.
11. Add water to the dry soil 2 days before plantlet transfer ensured that soil and water could properly mix.

References

1. Kellogg EA (2015) *Brachypodium distachyon* as a genetic model system. Annu Rev Genet 49:1–20
2. Vogel JP, Tuna M, Budak H et al (2009) Development of SSR markers and analysis of diversity in Turkish populations of *Brachypodium distachyon*. BMC Plant Biol 9:88
3. Skirycz A, Vandenbroucke K, Clauw P et al (2011) Survival and growth of *Arabidopsis* plants given limited water are not equal. Nat Biotechnol 29:212–214
4. Ariel F, Romero-Barrios N, Jégu T et al (2015) Battles and hijacks: noncoding transcription in plants. Trends Plant Sci 20:362–371
5. Verelst W, Bertolini E, De Bodt S et al (2013) Molecular and physiological analysis of growth-limiting drought stress in brachypodium distachyon leaves. Mol Plant 6:311–322
6. Axtell MJ (2013) ShortStack: comprehensive annotation and quantification of small RNA genes. RNA 19:740–751
7. Fahlgren N, Carrington JC (2010) miRNA target prediction in plants. Methods Mol Biol 592:51–57
8. Wu H-J, Wang Z-M, Wang M, Wang X-J (2013) Widespread long noncoding RNAs as endogenous target mimics for MicroRNAs in plants. Plant Physiol 161(4):1875–1884

Chapter 3

Brachypodium distachyon Long Noncoding RNAs: Genome-Wide Identification and Expression Analysis

Concetta De Quattro, Erica Mica, Mario Enrico Pè, and Edoardo Bertolini

Abstract

Recent advances in high throughput sequencing technology have revealed a pervasive and complex transcriptional activity of all eukaryotic genomes and have allowed the identification and characterization of several classes of noncoding RNAs (ncRNAs) with key roles in various biological processes. Among ncRNAs, long ncRNAs (lncRNAs) are transcripts typically longer than 200 nucleotides whose members tend to be expressed at low levels, show a lack of phylogenetic conservation and exhibit tissue-specific, cell-specific, or stress-responsive expression profiles.

Although a large set of lncRNAs has been identified both in animal and plant systems, the regulatory roles of lncRNAs are only beginning to be recognized and the molecular basis of lncRNA mediated gene regulation remains largely unexplored, particularly in plants.

Here, we describe an efficient methodology to identify long noncoding RNAs using next-generation sequencing data.

Key words Noncoding genome, Long noncoding RNAs, Next generation sequencing, RNA-seq, *Brachypodium distachyon*

1 Introduction

New high-throughput sequencing technology has revolutionized life science, allowing researchers to produce a large set of next generation sequencing (NGS) data in the last decade [1]. Transcriptomics has provided a clear evidence of a pervasive eukaryotic genome transcription, allowing the identification of new transcripts belonging to coding and noncoding RNAs (ncRNAs). Noncoding RNA molecules comprise two classes of transcripts: (a) structural ncRNAs, mainly ribosomal RNAs, transfer RNAs, small nuclear RNAs and small nucleolar RNAs and (b) regulatory ncRNAs, i.e., small RNAs (microRNAs and small interfering RNAs with a sequence length ranging between 21 and 24 nt) and long noncoding RNAs (lncRNAs), which are generally defined as transcripts longer than 200 nt, often capped, spliced, and polyadenylated

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_3,

[2]. This emerging latter class is increasingly recognized as a functional regulatory component in eukaryotic gene regulation, but few clear examples of gene regulation mediated by lncRNAs have been described in plants so far. For example, the Arabidopsis *COLDAIR* and *COOLAIR* regulate the expression of *FLOWERING LOCUS C* (*FLC*) during vernalization [3, 4], while the lncRNA *LDMAR* regulates the photoperiod-sensitive male sterility in rice [5]. In addition, target mimicry activity, which acts by attenuating the posttranscriptional repression of microRNA target genes, was described in Arabidopsis, mediated by the lncRNA *INDUCED BY PHOSPHATE STARVATION* acting as a decoy of miR399, which then is unable to interact with its target gene *PHO2* [6].

To date in plant species lncRNA annotation projects have been completed in *Arabidopsis thaliana* [7], *Oryza sativa* [8], *Zea mays* [9], and *Solanum lycopersicum* [10]. Important features of lncRNAs, such as lack of sequence conservation, tissue specificity and low expression levels make the annotation process difficult and dependent on the methodology used.

Here we present a computational method, based on bioinformatics tools freely available on the web, which allows to efficiently annotate lncRNAs of *B. distachyon* and other plant species from RNA-seq data and to estimate their expression profiles. Our in silico approach (shown in Fig. 1) starts with a raw FASTQ file generated by Illumina sequencers and ends up with a list of annotated noncoding transcripts and their expression profile values.

2 Materials

2.1 Software

1. 64-bit computer running Linux with at least 16 GB of RAM (*see* **Note 1**).
2. FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
3. Cutadapt (http://cutadapt.readthedocs.org/en/stable/guide.html).
4. ERNE-FILTER (http://erne.sourceforge.net/).
5. Bowtie software (http://bowtie-bio.sourceforge.net/index.shtml).
6. TopHat software (https://ccb.jhu.edu/software/tophat/index.shtml).
7. Cufflinks software (http://co1le-trapnell-lab.github.io/cufflinks/).
8. BLAST tools (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download).
9. CPC software (http://cpc.cbi.pku.edu.cn/).

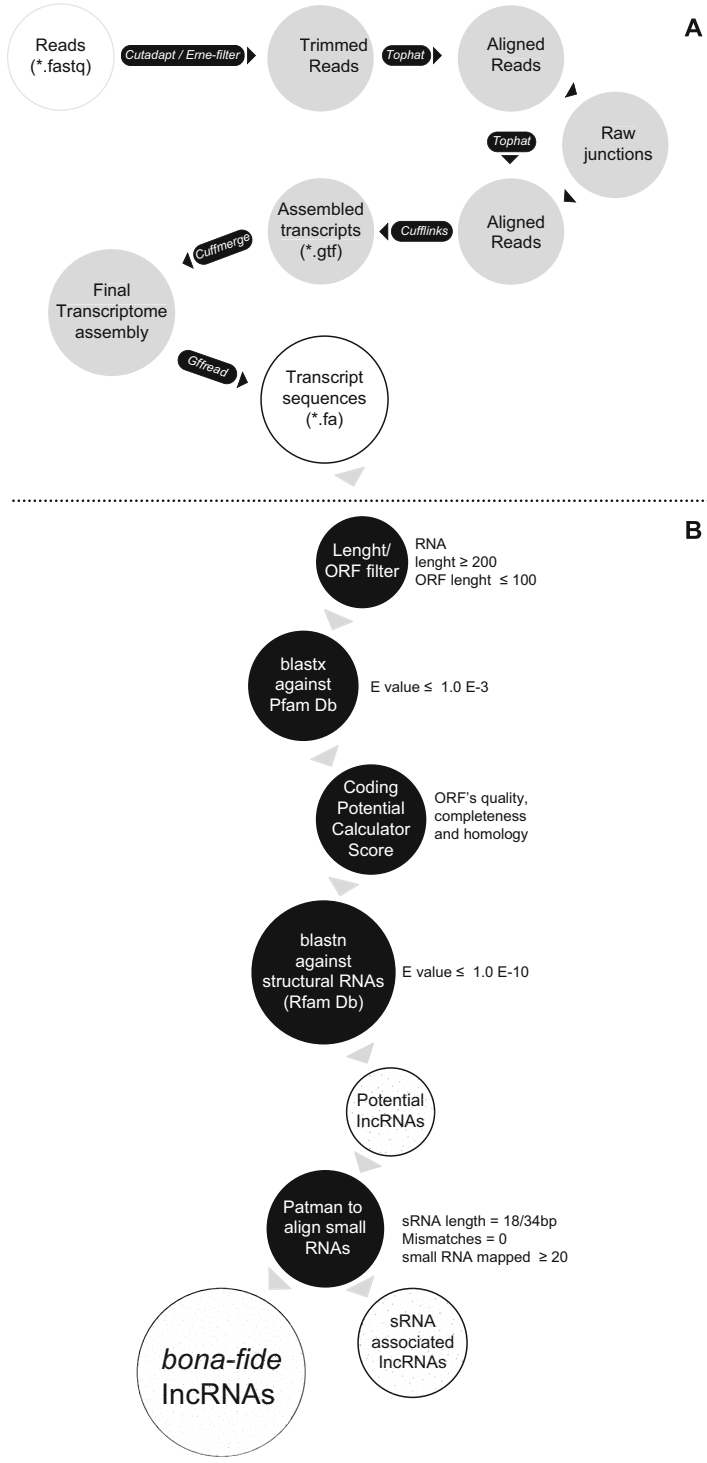

Fig. 1 Pipeline for the discovering of lncRNAs (**A**) The analysis of lncRNAs starts from the RNA-seq libraries available, either proprietary or available from databases. Libraries must be subjected to quality control: adapters and low quality reads are filtered through a two steps process involving Cutadapt and Erne-filter.

10. PatMaN software (http://bioinf.eva.mpg.de/patman).
11. Emboss Infoseq (http://emboss.bioinformatics.nl/cgi-bin/emboss/help/infoseq).
12. HTSeq-count (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html).
13. R software (http://www.r-project.org/).
14. edgeR Bioconductor R package (http://bioconductor.org/packages/release/bioc/html/edgeR.html).

2.2 Data Required to Perform the Analysis

1. mRNA-seq libraries in FASTQ format (*see* **Note 2**).
2. *Brachypodium distachyon* reference genome and annotation (http://phytozome.jgi.doe.gov/pz/portal.html) (*see* **Note 3**).
3. Pfam protein database latest version (http://pfam.xfam.org) (*see* **Note 4**).
4. *Brachypodium distachyon* structural RNAs from Rfam database latest version (http://rfam.xfam.org/) (*see* **Note 5**).
5. *Brachypodium distachyon* small RNA libraries in FASTA format (*see* **Note 6**).

3 Methods

This protocol allows the identification of long noncoding RNAs starting from RNA-Seq data (Fig. 1). The protocol is divided into three parts: (i) RNA-seq libraries management and analysis, (ii) identification of lncRNAs, and (iii) lncRNAs expression profiles analysis.

3.1 RNA-Seq Libraries Management and Analysis

This protocol starts from one or more raw RNA-seq libraries. First, the quality of each raw library is analyzed and sequencing adapters and bad quality reads are removed. Then the Tuxedo pipeline is performed, starting from TopHat, Cufflinks and Cuffmerge

Fig. 1 (continued) Trimmed reads are then align two times against to the reference genome using TopHat2. The second alignment is done to improve the reads mapping around spliced sites using information of junction sites generated from the first alignment. Mapping file (.bam) are processed using Cufflinks, which assembles the whole transcripts set for each sample. The assembled transcriptomes are then merged in a unique reference transcriptome with Cuffmerge and transcript sequences are extracted with Gffread. (**B**) Transcript sequences are subjected to six consecutive filters to identify long noncoding RNAs. The filters take into account the main features characterizing lncRNAs: (i) length $\geq$ 200 bp; (ii) ORF $\leq$100 amino acids; (iii) no homology with known protein domain; (iv) a low coding potential; (v) no homology with structural RNAs; (vi) no association with small RNAs. sRNAs associated with lncRNAs are defined as long noncoding RNA with sRNA reads mapped on their sequences. Vice versa sequences passing all the filters are classified as *bona fide* lncRNAs. Arrows indicate transcripts that pass the filters

(Fig. 1a). A detailed description of the Tuxedo workflow is explained in Trapnell et al. [11]. In order to use the information of splice sites derived from all samples, two iterations of TopHat are executed as proposed by Cabili et al. [12].

3.1.1 Quality Control and Filtering mRNA-seq Libraries

1. Collect proprietary or public RNA-seq data.
2. Check the quality of each library with FastQC tool. Check FastQC output with particular attention to the output test: (a) *Per base sequence quality*, (b) *Per base quality score*, (c) *Over-represented sequences*, and (d) *Adapter content*. The FastQC output helps to set program option in **step 3** and **step 4**.

```
fastqc name_library.fastq
```

3. Remove adapters with Cutadapt [13] choosing the most suitable parameters to trim your RNA-seq library. Use the options *-b (-anywhere = ADAPTER)* to trim the adapter sequence and the option *-O (-overlap = LENGTH)* to set the minimum overlap length between the adapter and the reads. In addition, fix the minimum reads length to be retained after adapter trimming through the parameter *-m* (*-minimum-length*).

```
cutadapt -b -O -m name_library.fastq -o name_library_trimmed.
fastq
```

4. Filter low quality reads using ERNE-FILTER [14]. To exclude low quality reads set the minimum mean PHRED value to 25 with the option *-min-mean-phred-quality*. In addition, to reduce multi hits mapping set the minimum sequence length after trimming with the option *-min-size. erne-filter -query1 name_library_trimmed.fastq -min-mean-phred-quality-min-size-output-prefix name_library_HighQuality*

5. Check again with FastQC the quality of the library resulting from **step 3** and **4** to be sure that reads do not contain adapters or over-represented sequences.

```
fastqc name_library_HighQuality.fastq.
```

3.1.2 Alignment to the Reference Genome

1. Download *Brachypodium distachyon* genome assembly and genome annotation.
2. Build the reference genome index for bowtie2 using bowtie2-build command.

```
bowtie2-build genome_reference.fa Bdistachyon_283
```

3. Perform two iterations of reads alignment using TopHat:

(a) First align RNA-seq reads of each sample independently to the reference genome without providing the GFF3 annotation file (*see* **Note 8**). Set the minimum intron length to 5 and the maximum intron length to 60,000 using the *-i* (*-min-intron-length*) and `--I` (`--max-intron-length`) options.

```
tophat2 -i -I -o name_library_first_alignment Bdistachyon_283 name_library_HighQuality.fastq
```

(b) Create a pooled splice-sites database (merged_junctions.bed) combining the junctions.bed files obtained from the first run of TopHat for each RNA-seq library. Use the command *cat* to concatenate the junctions.bed files.

```
cat junctions_tophat1.bed junctions_tophat2.bed >merged_junctions.bed
```

(c) Convert the pooled merged**.**junctions.bed file into a junction file (.juncs format) using the command *bed_to_juncs* available in the TopHat suite. Order and remove the duplicates within the merged**.**junctions.bed file using the bash commands *sort* and remove the duplicates with the option *-u* (*-unique*) (*see* **Note 9**).

```
bed_to_juncs < merged_junctions.bed | sort -k 1,4 -u| sort -k 1,1 >merged_junctions.juncs
```

(d) Realign each RNA-seq sample using the option *-j (-raw-juncs*) to provide the pooled junctions file produced in the previous step and the option *-no-novel-juncs* which allows to consider only reads across junctions indicated in the junctions file. In this second alignment, the reference gene model annotations (.GFF3) is provided with the option `-G`.

```
tophat2 -i -I -jmerged_junctions.juncs-no-novel-juncs -G Bdistachyon_283_v2.1.gene_exons.gff3 -o name_library_second_alignment Bdistachyon_283 name_library_HighQuality.fastq
```

3.1.3 Transcriptome Reconstruction

1. Reconstruct de novo the transcripts independently for each RNA-seq library using Cufflinks (*see* **Note 10**). In particular, supply the reference gene annotation (GFF3 file) using the parameter *-g* which guides the assembly and allows the assembly of novel transcripts and isoforms.

```
cufflinks -o name_library_assembly -I -g Bdistachyon_283_v2.1.gene_exons.gff3 name_library_second_alignment/accepted_hits.bam
```

2. Create a text file named *bd21_assembled_transcripts.txt* containing the full path of the assembly GTF file for each sample produced by Cufflinks.
3. Create a nonredundant final transcriptome GTF assembly for all the GTF assemblies present in *bd21_assembled_transcripts.txt* file using Cuffmerge (*see* **Note 10**).

```
cuffmerge -g Bdistachyon_283_v2.1.gene_exons.gff3 -s Bdistachyon_283.fa -o merged_bd21 bd21_assembled_transcripts.txt
```

4. Extract transcripts sequences in FASTA format using gffread utility with the options *-F*, *-Z*, *-E*, *-w* (*see* **Note 10**).

```
gffread merged_bd21/merged.gtf -g Bdistachyon_283.fa -F -o bd21_newGFFread.gff -Z -w bd21_transcripts.fa -W -E
```

3.2 LncRNAs Identification

This section describes the workflow to identify lncRNA molecules (Fig. 1B). Six different filters are applied to the unique set of transcripts sequences: length selection, Open Reading Frame length selection, known protein domain filter, Coding protein potential filter, structural ncRNA filter, and smallRNA filter. Sequences passing all the entire set of filters are defined as bona fide lncRNAs.

3.2.1 Data Preparation

1. Download protein sequences in FASTA format from Pfam database.
2. Create a Brachypodium structural RNA database in FASTA format downloading sequences from Rfam database.
3. Create a Brachypodium smallRNA database in FASTA format.

3.2.2 Identification of lncRNAs

1. Discard transcripts shorter than 200 bp.
2. Discard transcripts with an open reading frame greater than 100 amino acid since most annotated lncRNAs have short ORF, even if a number of well characterized human lncRNAs were reported to have ORFs of >100 codons [15].
3. Eliminate all transcripts which contain a known protein domains aligning the transcript against a Pfam protein database by using BLASTX [16]. Set the E-value of the alignment against the protein database equal to 0.001.
4. Assess the protein-coding potential using the Coding Potential Calculator (CPC) with default parameters [17] (*see* **Note 11**). Retain transcripts that do not pass the CPC filter.
5. Exclude noncoding structural molecules aligning transcripts to the structural ncRNAs sequences downloaded from Rfam database using BLASTN (P < 1.0E-10). Retain only those sequences that do not have similarity with structural ncRNAs. Sequences passing this filter are classified as potential lncRNAs.
6. Identify small RNAs associated with lncRNAs by aligning Bd small RNA sequences with zero mismatches to the potential lncRNAs using PatMaN [18] (*see* **Note 12**). LncRNAs with sequence homology to small RNA are classified as sRNA associated lncRNAs.
7. Transcripts passing all the six filters are then considered *bona fide* lncRNAs.

3.3 Expression Analysis of LncRNAs

This section describes how to identify the expression profiles of bona fide lncRNAs. The analysis is carried out into two successive steps: (a) creation of a matrix containing reads counts, (b) evaluation of the lncRNAs expression profiles normalized in Reads Per Kilobase per Million mapped reads (RPKM).

3.3.1 Creation of a Count Matrix

1. Calculate *bona fide* lncRNAs length by using Emboss Infoseq with the options *-name* and `-length`.

```
infoseq bonafide_lncbd21 -outfile bonafide_lncbd21_info \
-only -name -length
```

2. Count the number of mapped reads overlapping transcript sequences in each sample using the program HTSeq-count [19]. Perform this analysis using the sequence alignment file (.BAM/.SAM) generated for each sample and the gene model annotation file (.GFF) obtained from the utility gffread. We recommend to discard reads mapping multiple times in the genome to exclude false positives when differential expression analysis is performed. The output is a table where rows contain the transcript name and column the counts of mapped reads.

3. Create a count matrix merging together the count file of all sample analyzed. Rows in the count matrix will contain the transcripts names, while columns the raw counts for each experiment.

3.3.2 lncRNA Expression Analysis

Calculate the expression level of bona fide lncRNAs in RPKM using edgeR Bioconductor R package [20]. RPKM normalizes the expression based on library size and transcripts length. To normalize lncRNAs expression, perform the following steps:

1. Create a vector of lncRNAs lengths importing the file obtained with Emboss infoseq into R environment.
2. Import the count matrix into the R environment.
3. Check the mode of the two R object: count matrix and vector length. Both need to be numeric, otherwise convert them to numeric.
4. Compute reads per kilobase per million (RPKM) values using the `rpkm` function in edgeR given as inputs the count matrix (x) and the vector with lncRNAs length. The rows of the count matrix and the vector of length have the same order.
5. Extract the RPKM values related to lncRNAs with the R function *merge* using as common field the headers name of *bona fide* lncRNAs.

4 Notes

1. Most of the programs and commands given in this protocol can be run at the UNIX shell prompt. The protocol also includes small sections of code runnable in the R statistical computing environment (Section 3.3.2.).
2. Public and proprietary RNA-seq libraries can be used to identify lncRNAs. Public RNA-seq libraries can be downloaded from different NGS database, in particular from the Sequence Read Archive (SRA) at the National Centre for Biotechnology Information (NCBI) [21]. This protocol has been tested on 11 RNA-seq libraries downloaded from http://www.ncbi.nlm.nih.gov/sra (SRA: SRP008505).
3. This method was tested using *Brachypodium distachyon* Bd21 genome assembly version 2.0 and gene models version 2.1 released by the US Department of Energy (DOE) Joint Genome Institute (JGI) [22].
4. Pfam database is a protein database which includes 16,230 protein families in the version 28.0 [23].
5. *Brachypodium* genomic and plastidial tRNA, rRNA, small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) were downloaded from Rfam database version 12.0 [24].

6. *Brachypodium* smallRNA libraries can be downloaded online from the SRA database (reads are in raw format and must be cleaned from sequencing adapter before being used) and from the Plant MPSS database (mpss.udel.edu) produced by the Meyers' lab at the University of Delaware [25].
7. RNA-seq libraries downloaded from SRA database are in .sra format, hence in order to perform the computational pipeline, libraries must be converted into FASTQ format.
8. TopHat is a spliced aligner for RNA-seq experiment. Before aligning the reads to the reference genome, it is necessary to choose the appropriate parameters according to mRNA-seq libraries features. A full description of all the parameters is provided in the TopHat manual available on the website (https://ccb.jhu.edu/software/tophat/manual.shtml).
   HISAT2 is the recent successor of TopHat (http://ccb.jhu.edu/software/hisat2/) and in this methods can be adopted in place of TopHat in the mapping step.
9. The command *bed_to_juncs* is included in the TopHat software and it converts the file junctions.bed into a junctions file which can be provided to TopHat with the option `-j`.
10. Cufflinks is a suite of tools that can be used to analyze RNA-seq experiments. Cufflinks suite programs used in the analysis described above are Cufflinks, Cuffmerge and gffread utility. The parameters to be considered during the analysis are chosen based on the characteristics of the RNA-seq libraries and the purpose of the analysis. The detailed description of each parameter is given in the Cufflinks manual which is available on the web site (http://cole-trapnell-lab.github.io/cufflinks/manual/). StringTie is the successor of Cufflinks (https://ccb.jhu.edu/software/stringtie/) and can be adopted in the transcriptome assembly in place of Cufflinks.
11. Coding Potential Calculator (CPC) discriminates protein coding from noncoding genes using a sequence homology based approach and a Support Vector Machine classifier to assess the protein coding potential of a transcript. It uses a comprehensive and updated UniProt database for training the algorithm. Recently the updated version of the program (CPC2) has been released.
12. PatMaN allows to align short smallRNA reads in FASTA format against the desired genome/transcriptome file in FASTA format. The specific parameters to be set for this analysis are: 18–34 nt sequences length, 0 mismatch between short reads and genome/transcriptome sequences. Transcripts with twenty or more different small RNA reads mapped on their sequences are marked as lncRNAs associated with small RNAs.

Acknowledgment

This work was supported by the International Doctoral Programme in Agrobiodiversity of Scuola Superiore Sant'Anna (http://www.santannapisa.it/).

References

1. Martin LBB, Fei Z, Giovannoni JJ et al (2013) Catalyzing plant science research with RNA-seq. Front Plant Sci. doi:10.3389/fpls.2013.00066
2. Kim ED, Sung S (2012) Long noncoding RNA: unveiling hidden layer of gene regulatory networks. Trends Plant Sci 17:16–21
3. Heo JB, Sung S (2011) Vernalization-mediated epigenetic silencing by a long Intronic noncoding RNA. Science 331:76–79
4. Swiezewski S, Liu F, Magusin A et al (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. Nature 462:799–802
5. Ding J, Lu Q, Ouyang Y et al (2012) A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proc Natl Acad Sci 109:2654–2659
6. Franco-Zorrilla JM, Valli A, Todesco M et al (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. Nat Genet 39:1033–1037
7. Wang H, Chung PJ, Liu J et al (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in Arabidopsis. Genome Res 24:444–453
8. Zhang YC, Liao JY, Li ZY et al (2014) Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol 15:512
9. Li L, Eichten SR, Shimizu R et al (2014) Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol 15:R40
10. Zhu B, Yang Y, Li R et al (2015) RNA sequencing and functional analysis implicate the regulatory role of long non-coding RNAs in tomato fruit ripening. J Exp Bot 66:4483–4495
11. Trapnell C, Roberts A, Goff L et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nat. Protoc* 7:562–578
12. Cabili MN, Trapnell C, Goff L et al (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915–1927
13. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17:10–12
14. Del Fabbro C, Scalabrin S, Morgante M et al (2013) An extensive evaluation of read trimming effects on Illumina NGS data analysis. PLoS One. doi:10.1371/journal.pone.0085024
15. Housman G., Ulitsky I. (2015) Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of translation of long noncoding RNAs. BBA-Gene Regul Mech http://dxdoiorg/101016/jbbagrm201507017 Available online 8 August 2015
16. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. BMC Bioinform 10:421
17. Kong L, Zhang Y, Ye ZQ et al (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35:W345–W349
18. Prüfer K, Stenzel U, Dannemann M et al (2008) PatMaN: rapid alignment of short sequences to large databases. Bioinformatics 24:1530–1531
19. Anders S, Pyl PT, Huber W (2015) HTSeq—a python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169
20. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140
21. Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database Collaboration (2012) The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 40:D54–D56

22. International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768
23. Finn RD, Bateman A, Clements J et al (2014) The Pfam protein families database. Nucleic Acids Res. doi:10.1093/nar/gkt1223
24. Nawrocki EP, Burge SW, Bateman A et al (2014) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. doi:10.1093/nar/gku1063
25. Nakano M, Nobuta K, Vemaraju K et al (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res 34: D731–D735

Chapter 4

A Highly Efficient and Reproducible *Fusarium* spp. Inoculation Method for *Brachypodium distachyon*

Anuj Rana, Aneesh Karunakaran, Timothy L. Fitzgerald, Rosalie Sabburg, Elizabeth A.B. Aitken, Robert J. Henry, Jonathan J. Powell, and Kemal Kazan

Abstract

Fusarium spp. are devastating fungal pathogens which cause significant losses in many cereal crops like wheat, maize, and barley. Genetic improvement of disease resistance requires an improved understanding of defense-associated processes operating in the host in response to an attack by *Fusarium* spp. *Brachypodium distachyon* is emerging as a model where host–cereal-infecting pathogen interactions can be studied conveniently. However, this requires developing an efficient infection assay that facilitates quick screening of germplasm (e.g., mutant lines). Here, we provide an efficient and reproducible *Fusarium* infection assay for *Brachypodium*. We believe this method will help further develop *Brachypodium* as a model for genetic improvement of disease resistance in cereals against *Fusarium* pathogens.

Key words *Fusarium* spp., Defence marker genes, *Brachypodium distachyon*, Head blight, Fungal biomass

1 Introduction

Fungal pathogens are among the leading biological stress factors causing substantial crop losses worldwide. *Fusarium* pathogens are a diverse group of fungi that cause diseases on hundreds of plant species, including major cereal crops. For instance, Fusarium head blight (FHB) and crown rot are the major fungal diseases that cause significant losses in cereal (wheat, barley and maize) production. *Fusarium graminearum* is the primary causative agent of the FHB disease. This pathogen also produces trichothecene mycotoxins during infection of cereal heads [1, 2]. Deoxynivalenol (DON) and nivalenol (NIV) are major toxins produced during the fungal infection [3] and can contaminate food and feed produced from infected plant material. They can be harmful for humans and animals if contaminated products are consumed.

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_4,

The related *Fusarium* species *Fusarium pseudograminearum* infects the stem base of wheat and barley plants causing the Fusarium crown rot (FCR) disease [4], FCR is of economic significance in several countries, including China, Turkey, United States and Australia [5]. Resistance to FHB and FCR is highly quantitative in nature [6] and this makes breeding efforts difficult. Consequently, molecular characterization of host defense is required to develop tractable sources of resistance for both diseases although this remains technically difficult due to the complexity of the bread wheat genome.

Brachypodium distachyon is emerging as model for monocot cereals, having diverged from Triticeae between 35 and 40 million years ago [7]. *B. distachyon* is more closely related to the Triticeae family than rice or other sequenced grasses such as sorghum and maize [8]. *B. distachyon* has a very small genome (272 Mbp) with five chromosomes that show close synteny to wheat [9]. *Brachypodium* is also easily transformable and has extensive and publically available genetic and genomic resources (e.g., ecotypes, mutants, T-DNA insertion lines etc. [10]). To be able to use *Brachypodium* as a model pathosystem, it is important to show its susceptibility to cereal pathogens. Indeed, Fusarium infection systems have been developed for *Brachypodium* in recent years [11, 12]. Here, we provide efficient, easy, and reproducible FHB and FCR infection assays for *Brachypodium*, which can be used to understand potential mechanisms involved in host resistance against these pathogens.

2 Materials

1. *Brachypodium distachyon* (Bd21-3) seeds.
2. Sterilized distilled water.
3. *Fusarium graminearum* and *Fusarium pseudograminearum* isolates.
4. Erlenmeyer flask (250 mL).
5. Potato dextrose broth (powder) (PDB) (Difco, Sparks, MD, USA).
6. Campbell's V8 Juice (Campbells Australia, Shepparton, Victoria): Add 200 mL of V8 juice with distilled water and prepare the final volume to 1000 mL. Autoclave the medium for use.
7. Carboxy methyl-cellulose (CMC) medium: 15 g carboxy methyl-cellulose (Sigma), 1 g NH_4N0_3, 1 g KH_2PO_4 monobasic, 0.5 gm $MgSO4.7H_2O$, 1 g Yeast Extract, and 1000 mL distilled water. Add all ingredients into gently boiling water and mix vigorously using a magnetic stirrer until all CMC dissolves. Autoclave the medium before use.

8. Agar medium (0.8%): Weigh 8.0 g of Bacto agar and dissolve it in 1000 mL distilled water. Autoclave the medium before use.
9. Technical Agar (Bacto, Mt. Pritchard, Australia).
10. Whatman™ filter paper No. 3 [12.5 cm].
11. Tween® 20 (Sigma, St Louis, MO).
12. Glass rod.
13. Sterilized disposable needles.
14. Sterilized surgical blades.
15. Gloves.
16. Cork borer (Stainless steel).
17. Hemocytometer.
18. Light/compound microscope.
19. Laminar air flow.
20. Growth cabinet.
21. Rotary shaker.
22. NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, DE, USA).
23. Large round petri plates [14 cm] (Corning).
24. Plastic pots (10 cm diameter ANOVA pot).
25. Potting mix (Searles Peat 80 Plus® premium potting mix).
26. Polypropylene tubes [15 and 50 mL].
27. Microfuge tubes.
28. Glycerol (40%).
29. Dried and clean cracked corn kernels.
30. Parafilm or plate wrap.
31. Benchtop centrifuge.
32. 3% available hypochlorite solution.
33. Ethanol 70%.
34. Nuclease free water.
35. Ambion™ RNase away decontamination reagent (Thermo Fisher Scientific, Carlsbad, CA, USA).
36. MicroAmp® Optical 384-Well Reaction Plate with Barcode (Thermo Fisher Scientific, Carlsbad, CA, USA).
37. DNeasy® extraction kit (Qiagen, Hilden, Germany).
38. RNeasy® plant mini kit (Qiagen, Hilden, Germany).
39. First strand cDNA synthesis kit (Invitrogen, Carlsbad, CA, USA).
40. SYBR® Green master mix (2×) (Applied Biosystems, Warrington, UK).

41. 3 mm stainless steel beads (Treated with RNase away).
42. Retsch mixer mill MM400 (Retsch, GmbH, Haan, Germany).

3 Methods

3.1 Fusarium Head Blight Infection Assay

3.1.1 Fungal Culture Preparation

1. Inoculate ¼ strength PDA plates (6 g Potato Dextrose Broth and 16 g Technical Agar L^{-1} water) with a single hyphae of the fungal strain in laminar flow and incubate at 25 °C under fluorescent lighting for 5–7 days (*see* **Note 1**).
2. Cut out circular agar from fungal hyphae grown on the PDA plate by a sterilized cork borer. Use 7–8 circular plugs of fungal mycelium taken from the edge of a fully grown fungal culture plate and inoculate into 250 mL flask containing 70 mL V8 juice or the CMC broth medium (*see* **Note 2**).
3. Shake the inoculated flask at 150 rpm on a rotary shaker at room temperature for 5–7 days or the culture becomes cloudy.
4. Filter the uniformly grown fungal culture through miracloth and centrifuge the filtrate at 2000 × *g* at 4 °C for 10 min. After centrifugation, gently discard the supernatant and resuspend the pellet in 5 mL of sterilized distilled water and repeat **step 2** (*see* **Note 2**).
5. Dissolve the pellet in 1 mL sterile distilled water. Count and dilute the spores at a concentration of 1×10^6 spores/mL using a hemocytometer under light/compound microscope.
6. Store the fungal spores at 4 °C and use fresh spores for infection (*see* **Note 3**).

3.1.2 Plant Growth Conditions

1. Keep *B. distachyon* seeds for 1–3 weeks at 4 °C in petri plates containing a moist filter paper for vernalization (*see* **Note 4**).
2. Plant around ten germinated seeds in a plastic pot (8 × 8 × 10 cm) containing soil compost mixture and keep them in a controlled growth chamber at 25 °C day and 16 °C night and 18 h–6 h light–dark cycle with 60% relative humidity (*see* **Note 5**).
3. Grow the plants under the same conditions and watered every 2–3 days for 6 weeks up to the flowering stage (*see* **Note 5**).

3.1.3 Detached Leaf Inoculation

Cut out the leaf tissue with petiole from healthy plants 3 weeks after sowing and collect the samples in large sterile petri plates containing distilled water.

1. Place an individual leaf on a sterilized tissue paper and wound the leaf tissue from the adaxial site of the explant by gently pressing the glass rod onto the tissue under sterile conditions. Move the wounded explants on an agar plate (0.8% agar medium) by inserting the petiole into the agar.

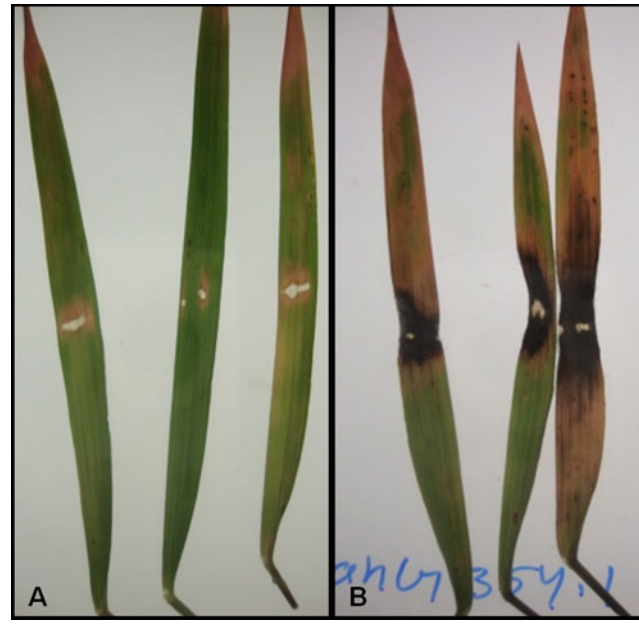

Fig. 1 Scanned image of *Brachypodium* leaves: (**a**) mock, (**b**) each leaf was inoculated with *Fusarium graminearum* spores (2×10^3 spores) suspension after wounding

2. Inoculate the leaf by pipetting 2 μL of spore suspension (10^6 spores/mL) amended with 0.05% Tween 20 on the wound site and incubate the explants in a growth chamber under the same conditions used previously for plant growth.
3. Scan the infected leaves with a flat-bed scanner and measure the necrotic leaf area by an image software (ImageJ v1.49) at different time intervals such as 3, 5 and 7 days post inoculation (Fig. 1).

3.1.4 Head Blight Inoculation of Intact Floral Tissue

1. Select at least ten healthy spikes per plant after 6 weeks of sowing and place the filter paper (~2 × 6 mm) between the second and the third spikelet from the bottom (Fig. 2).
2. Inoculate single spikelet of individual spikes by applying 5 μL of spore suspension (10^6 spores mL^{-1} amended with 0.05% Tween 20) directly onto the filter paper (*see* **Note 6**).
3. Observe floral disease symptoms (tissue browning and necrosis) and record the disease development at 3, 7, 9, 15, and 21 days post-inoculation. A disease index can be calculated by dividing the number of infected spikes to total number of inoculated spikes per plant. Score disease symptoms at different time points on a 0–5 scale, where 0 = no symptom, 1 = <20%, 2 = 21–40%, 3 = 41–60%, 4 = 61–80%, and 5 = 81–100% of spikelet with necrosis (Fig. 3).

Fig. 2 Brachypodium intact floral spikes inoculated with *Fusarium graminearum* spores (5×10^3 spores) suspension. *White arrows* indicate the position of the filter paper used in inoculation experiment

Fig. 3 Disease scoring for percent (%) infection: (**a**) 5 (81–100%) (**b**) 4 (61–80%) (**c**) 3 (41–60%) (**d**) 2 (21–40%) (**e**) 1 (<20%) (**f**) 0 (0%)

3.2 Fusarium Crown Rot Infection Assay

3.2.1 Fusarium Agar Plug Inoculum

1. Inoculate the center of V8 agar plates with a *Fusarium* isolate of interest in a laminar air flow.
2. Incubate the plates for 5–7 days under 12 h light/12 dark at 25 °C.
3. Under sterile conditions, use the wide end of a sterile 200 μL pipette tip to excise round plugs from the growing edge of the plate.

3.2.2 Fusarium Root Rot Inoculation

1. Peel lemma from *Brachypodium* seeds using forceps.
2. Surface sterilize the seeds using a 3% hypochlorite solution followed by 70% ethanol and rinse 3–4 times with sterile water to remove residual hypochlorite and ethanol.
3. Place 10–15 seeds on two autoclaved Whatman No. 3 filter papers placed inside a 14 cm petri dish and keep the plates at 4 °C for 1 week to stratify the seeds.
4. Once *Brachypodium* seedlings have germinated and produced a primary root of 3–4 cm in length, place the mycelium side of the agar plugs (as described above) down on the root approximately 2 cm away from the seed.
5. Seal the plates with Parafilm and incubate them in a growth cabinet under 16 h light/8 h dark at 22 °C.
6. Observe disease progression at different time points of interest; 7 and 14 days post inoculation time points are recommended (Fig. 4).

3.3 Quantitative Techniques for Fungal Infection Using qRT-PCR

3.3.1 Tissue Harvesting

1. Produce *Fusarium* infected plant tissue using one of the procedures described above (Subheadings 3.1 and 3.2).
2. Harvest the tissue of interest at relevant time points and collect it into nuclease-free collection tubes containing 3 mm stainless steel beads. Snap freeze the tissue immediately in liquid nitrogen and keep at −80 °C until tissue disruption.
3. Grind to a fine powder in a MM400 ball mill.

Fig. 4 *Brachypodium distachyon* seedlings infected with *Fusarium pseudograminearum* using root rot inoculation method

3.3.2 Preparation of Standards

1. Grow a *Fusarium* isolate on media plates as described above (Subheading 3.2.1).
2. Scrape mycelium off the surface of plates and place it into polypropylene tubes. Freeze-dry material.
3. Harvest disease-free *Brachypodium* tissue and snap-freeze in liquid nitrogen. Freeze-dry material.
4. Disrupt freeze-dried material (*Fusarium* mycelium and disease-free *Brachypodium*) to a fine powder, by adding 3 mm stainless steel beads to each tube and grinding using a Retsch mixer mill.
5. To prepare five standards with a tenfold dilution series of 10–0.001 mg g^{-1}; first, make a stock standard (Standard 1) by mixing 10 mg of ground fungal mycelia with 1 g of ground disease-free *Brachypodium* tissue. Mix thoroughly by vortexing.
6. Prepare Standards 1–4 (1–0.001 mg g^{-1}) as follows, mixing each standard thoroughly by vortexing, before making the next:

   Standard 2. Add 100 mg of Standard 1 to 900 mg disease-free *Brachypodium* tissue.

   Standard 3. Add 100 mg of Standard 2 to 900 mg disease-free *Brachypodium* tissue. Standard 4. Add 100 mg of Standard 3 to 900 mg disease-free *Brachypodium* tissue.

   Standard 5. Add 100 mg of Standard 4 to 900 mg disease-free *Brachypodium* tissue.
7. Extract genomic DNA for qRT-PCR from 20 mg of each standard using a DNeasy® extraction kit as per the protocol.
8. After extraction, measure DNA concentration and quality using a NanoDrop spectrophotometer.
9. Each standard should have a final concentration of 5 ng μL^{-1}.
10. Store the extracted DNA from standards at −20 °C until needed.

3.4 Fungal Biomass Quantification Using Quantitative Polymerase Chain Reaction (qPCR)

1. For each sample extract genomic DNA from 20 mg of finely ground tissue using a DNeasy® extraction kit as per the protocol.
2. After DNA extraction, measure DNA concentration and quality using a NanoDrop spectrophotometer and dilute samples to 5 ng μL^{-1} with MilliQ water in a 96-well plate.
3. Prepare 10 μL reactions in microtiter plate (MicroAmp® Optical 384-Well Reaction Plate) containing 4 μL of genomic DNA (5 ng μL^{-1}), 5 μL of SYBR® Green master mix, and 1 μL of primer mix (forward and reverse primers) (*see* **Note 7**). Include all five standards and water blanks (MilliQ water) as negative

controls. Include three technical replicates for each biological sample, standard, and negative control.

4. Place the reaction plate (MicroAmp® Optical 384-Well Reaction Plate) into a thermocycler equipped with a fluorescence detector (Viia7™ Real Time PCR System) and run with the following cycling parameters for both *F. pseudograminearum* and *F. graminearum* primers: initial heating steps of 50 °C for 2 min followed by 95 °C for 10 min prior to 40 cycling steps consisting of 15 s at 95 °C followed by 1 min at 60 °C. Gradient clines between set temperatures should be 1.6 °C/s except for the final step of the melt curve with a temperature gradient of 0.05 °C/s.
5. Analyze the results using Viia7™ software (Applied Biosystems). Perform quality checks to ensure amplification was not observed in negative controls. For each set of three technical replicates, ensure amplification curves are within one cycle of each other at the threshold. Calculate the mean of the technical replicates for use in subsequent calculations. Check melt curves and ensure only a single peak occurs across all samples for each specific primer pair.
6. Cycle threshold (cT) values of DNA from the standards were used to generate a standard curve. The regression curve should be linear and have an R^2 value >0.98.
7. To determine absolute *Fusarium* biomass solve for x in the regression equation for each sample (y = cT value) and express results as mg g^{-1} *Brachypodium* tissue.

3.5 RNA Extraction

1. Grind tissue samples into a fine powder using a tissue ruptor (Retsch ball mill or similar) set to 28 oscillations per second for 1 min and add extraction buffer directly onto the powder.
2. Follow RNeasy® Plant extraction kit protocol as per manufacturer's specification.
3. Perform on-column DNase treatment using Invitrogen RNase-free Dnase.
4. Store purified RNA at −80 °C until needed.

3.6 Quantitative Real-Time PCR

1. Perform cDNA synthesis using a reverse transcription protocol, e.g., SuperScript III first-strand synthesis kit according to manufacturers' specifications. Dilute cDNA to 1/100 with MilliQ water before using as a reaction template.
2. Design the primers for genes of interest using best practice guidelines [13]. Primers should be 18–22 bp in length; have a melting temperature between 58 and 60 °C, and a GC content of 40–60%; and amplify a product between 90 and

110 bp in length. Primer design software such as Primer3 [14] or PerlPrimer [15] are recommended.

3. BLAST the primer sequences against the *Brachypodium* genome reference to ensure the primers are specific for the target region of interest.
4. PCR reactions should be comprised of 5 μL SYBR green PCR mix, 1 μL of 3 μM working primer stock, and 4 μL of diluted cDNA. A minimum of three reactions for each sample–primer combination should be performed as technical replicates and negative controls should be incorporated to check for amplification in the absence of a template.
5. For each biological sample, set up reactions with reference gene and marker gene primer pairs.
6. Place the reaction plate (MicroAmp® Optical 384-Well Reaction Plate) into a thermocycler equipped with a fluorescence detector (Viia7™ Real Time PCR System) and run with the following cycling parameters: initial heating steps of 50 °C for 2 min followed by 95 °C for 10 min prior to 40 cycling steps consisting of 15 s at 95 °C followed by 1 min at 60 °C. Gradient clines between set temperatures should be 1.6 °C/s except for the final step of the melt curve with a temperature gradient of 0.05 °C/s.
7. Analyze the results using Viia7™ software (Applied Biosystems). Perform quality checks to ensure amplification was not observed in negative controls. For each set of three technical replicates, ensure that amplification curves are within one cycle of each other at the threshold. Calculate the mean of the technical replicates for use in subsequent calculations. Check melt curves and ensure only a single peak occurs across all samples for each specific primer pair.
8. For each biological sample, normalize cT values for each marker gene against cT values for the reference gene (*see* **Note 8**).
9. Calculate relative expression values (ΔΔcT or Rn values) for each biological sample–marker gene combination. Calculate mean relative expression values and compare to determine induction of defense gene during infection.

3.7 Conclusions

FHB and FCR infection methods described in this chapter are very efficient, reproducible and easy to perform in *Brachypodium*. These methods provide quick results for screening the susceptible and resistance plant lines and can be useful to employ in studies toward better understanding the role of genes involved at different stages of infection. The given protocols can be used for phenotypic and functional analysis of novel genes during plant-pathogen interaction in other cereal crops.

4 Notes

1. Carefully pick up the single growing hyphae with sterilized needle and inoculate it onto the center of the PDA plates.
2. Use a sterilized cork borer to cut out uniform holes in the freshly growing PDA plates or alternatively use a sterilized surgical blade to prepare uniform squares of agar in the plate.
3. Do not keep the spore dilution at 4 °C for too long. If possible, use freshly grown spores for infection.
4. Stratified the seeds before sowing for 1 week at 4 °C in the dark for detached leaf infection assay and 2–3 weeks to promote early flowering for head blight inoculation.
5. Perform the leaf and floral tissue infection assays simultaneously to obtain different measures of disease development with a minimum effort. Each treatment should require at least five individual replicates and each replicate with 5–10 head/leaf tissue. Do not forget to wet the soil and check the moisture level on alternate days.
6. Cut the Whatman™ filter paper No.1 in ~2 × 6 mm size pieces. Insert a single filter paper containing the inoculum gently between the spikelets. Infect nearly ten heads per plant and cover the infected plants with transparent polythene zip bags for the first 3 days post infection. Maintain the moisture by spraying sterilized distilled water into the polythene bags before covering the inoculated plants.
7. Primers targeting a *Fusarium* sequence in a species specific manner are required for accurate quantification. For *F. pseudograminearum* biomass quantification, the primer pair *Fp*tri3eF (5′-CAAGTTTGATCCAGGGTAATCC-3′) and *Fp*tri3eR (5′-GCTGTTTCTCTTAGT CTTCCTCA-3′) targetting3′) targeting the TRI3 gene is recommended [16]. For *F. graminearum* biomass quantification, the primer pair Tri6_10F (5′-TCTTTGTGAGCGGACGGGACTTTA-3′) and Tri6_4R (5′-ATC TCGCATGTTATCCACCCTGCT-3′) targeting the TRI6 gene is recommended [17]. Primer stocks had final concentration of 3 μM for both forward and reverse primers.
8. Ubiquitin conjugating enzyme 18 (Bradi4g00660) has been previously validated as a reference gene for qRT-PCR in *B. distachyon* [18]. Primers for classic defense marker genes which are responsive to *F. pseudograminearum* infection in leaf tissue are provided in Table 1.

Table 1
Primers for defense markers genes for use in qRT-PCR

Gene description	Gene ID (v1.2)	Forward primer	Reverse primer
Ubiquitin conjugating enzyme 18	Bradi4g00660	TTTACAGCAATGGCCACATC	AGACAGCATGGACAAGATGC
Pathogenesis Related 1	Bradi1g57590	TACCACCATGACGGGAATC	CACAAACAACACGAGCACAC
Pathogenesis Related 2	Bradi2g60490	AGCTTACAACCAGGGCTTGA	CGTTGAACATGGCAAAGATG
Pathogenesis Related 3	Bradi2g47210	AGCCATGACGTTATCACTGG	CCGTTGATGATGTTGGTGAT
Pathogenesis Related 4	Bradi4g14920	GACCTGGACTGGGACACG	TCACCACAGTCGACGAACTC

Acknowledgments

The authors are grateful to Grains Research and Development Corporation (GRDC), Australia for research funding under CSP00155. Anuj Rana is supported by an Endeavour fellowship from the Australian Government and University Grant Commission (UGC), New Delhi for postdoctoral research fellowship. Jonathan Powell is supported by the GRDC through provision of a graduate research scholarship (GRS10532).

References

1. Alexander NJ, McCormick SP, Waalwijk C, van der Lee T, Proctor RH (2011) The genetic basis for 3-ADON and 15-ADON trichothecene chemotypes in Fusarium. Fungal Genet Biol 48:485–495
2. Goswami RS, Kistler HC (2004) Heading for disaster: Fusarium graminearum on cereal crops. Mol Plant Pathol 5:515–525
3. Desjardins AE, Proctor RH (2007) Molecular biology of Fusarium mycotoxins. Int J Food Microbiol 119:47–50
4. Akinsanmi O, Backhouse D, Simpfendorfer S, Chakraborty S (2006) Genetic diversity of Australian Fusarium graminearum and F. Pseudograminearum. Plant Pathol 55:494–504
5. Liu C, Ogbonnaya FC (2015) Resistance to Fusarium crown rot in wheat and barley: a review. Plant Breed 134:365–372
6. Chakraborty S, Liu C, Mitter V, Scott J, Akinsanmi O, Ali S, Dill-Macky R, Nicol J, Backhouse D, Simpfendorfer S (2006) Pathogen population structure and epidemiology are keys to wheat crown rot and Fusarium head blight management. Australas Plant Pathol 35:643–655
7. Bossolini E, Wicker T, Knobel PA, Keller B (2007) Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. Plant J 49:704–717
8. Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci U S A 101:9903–9908

9. Huo N, Vogel JP, Lazo GR, You FM, Ma Y, McMahon S, Dvorak J, Anderson OD, Luo MC, Gu YQ (2009) Structural characterization of Brachypodium genome and its syntenic relationship with rice and wheat. Plant Mol Biol 70:47–61
10. Garvin DF, Gu YQ, Hasterok R, Hazen SP, Jenkins G, Mockler TC, Mur LAJ, Vogel JP (2008) Development of genetic and genomic research resources for *Brachypodium distachyon*, a new model system for grass crop research. Crop Sci 48:S69–S84
11. Fitzgerald TL, Powell JJ, Schneebeli K, Hsia MM, Gardiner DM, Bragg JN, McIntyre CL, Manners JM, Ayliffe M, Watt M, Vogel JP, Henry RJ, Kazan K (2015) Brachypodium as an emerging model for cereal-pathogen interactions. Ann Bot 115:717–731
12. Peraldi A, Beccari G, Steed A, Nicholson P (2011) *Brachypodium distachyon*: a new pathosystem to study Fusarium head blight and other Fusarium diseases of wheat. BMC Plant Biol 11:100
13. Udvardi MK, Czechowski T, Scheible W-R (2008) Eleven golden rules of quantitative RT-PCR. Plant Cell 20:1736–1737
14. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40:e115–e115
15. Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics 20:2471–2472
16. Khudhair M, Melloy P, Lorenz D, Obanor F, Aitken E, Datta S, Luck J, Fitzgerald G, Chakraborty S (2014) Fusarium crown rot under continuous cropping of susceptible and partially resistant wheat in microcosms at elevated CO2. Plant Pathol 63:1033–1043
17. Horevaj P, Milus E, Bluhm B (2011) A real-time qPCR assay to quantify Fusarium graminearum biomass in wheat kernels. J Appl Microbiol 111:396–406
18. Hong S-Y, Seo PJ, Yang M-S, Xiang F, Park C-M (2008) Exploring valid reference genes for gene expression studies in *Brachypodium distachyon* by real-time PCR. BMC Plant Biol 8:112

Chapter 5

Tissue Culture (Somatic Embryogenesis)-Induced *Tnt1* Retrotransposon-Based Mutagenesis in *Brachypodium distachyon*

Upinder S. Gill, Juan C. Serrani-Yarce, Hee-Kyung Lee, and Kirankumar S. Mysore

Abstract

Brachypodium distachyon is a model grass species for economically important cereal crops. Efforts are in progress to develop useful functional genomic resources in *Brachypodium*. A tobacco retrotransposon, *Tnt1*, has been used successfully in recent past to generate insertional mutagenesis in several dicot plant species. *Tnt1* retrotransposon replicates, transposes, and inserts at multiple random genomic locations in the plant genome. Transposition occurs only during somatic embryogenesis but not during seed transmission. We developed *Brachypodium* transgenic plants that can express the *Tnt1* element. Here, we describe an efficient tissue culture-based approach to generate *Tnt1* insertional mutant population using transgenic *Brachypodium* line expressing the *Tnt1* retrotransposon.

Key words *Brachypodium*, *Tnt1*, Insertional mutagenesis, Tissue culture

1 Introduction

Brachypodium distachyon (purple false brome) is an annual grass which is now established as a model monocot species to study grasses [1]. *Brachypodium* has been widely used as a model to study biotic stresses, abiotic stresses, plant growth and development, and cell wall biosynthesis [1, 2]. *Brachypodium* is a diploid with small genome (~272 Mb). The whole genome of *Brachypodium* was sequenced in 2010 [3]. *Brachypodium* is amenable to tissue culture and *Agrobacterium*-mediated transformation [4]. From the past few years, efforts are being made to generate a variety of genetic and genomic resources for this species. For example, to study gene functions in *Brachypodium*, T-DNA insertion mutant population has been generated [4, 5]. Since T-DNA causes on an average of only 1.5 insertions per line, achieving saturation or near saturation mutagenesis using T-DNA will be a daunting task.

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_5,

To address this issue, we used *Tnt1* retrotransposon-based mutagenesis in *Brachypodium*. The major advantage of using the *Tnt1* retrotransposon is that a large number of mutations can be generated in each plant because of the insertion of multiple copies of the *Tnt1* in the genome. This helps specially in forward genetics screening where more mutations can be studied using smaller population size. In addition, saturation or near saturation mutagenesis can be achieved using *Tnt1*. *Tnt1*-based insertional mutagenesis has been successfully used in the past for other plant species such as Arabidopsis, lettuce, potato, *Medicago truncatula*, and soybean [6–11]. *Tnt1* is unique because of its ability to transpose during somatic embryogenesis and sometimes during various biotic and abiotic stresses; however, transposition does not occur during seed-to-seed transmission under normal conditions [8, 12]. So far, use of *Tnt1*-based insertional mutagenesis has not been reported for any monocot species. We transformed *Brachypodium* via *Agrobacterium*-mediated transformation with the *Tnt1* retrotransposon and generated transgenic lines which express the *Tnt1* element (unpublished). Subsequent regeneration using the immature embryo of transgenic plants expressing *Tnt1* as an explant may lead to *Tnt1* transposition. Here, we describe a method for large-scale regeneration of *Brachypodium Tnt1* insertion mutant population using the transgenic Bd21-3 expressing the *Tnt1* element as a mother plant.

2 Materials

2.1 Plants and Plant Growth Materials

1. *Brachypodium* mother plant 7-1 (transgenic plant expressing *Tnt1*) seeds.
2. Plastic pots (4″).
3. Synthetic soil mixture (Metro-Mix 830, Sun Gro Horticulture, Agawam, MA, USA).
4. Linsmaier and Skoog basal medium (LSP03, Caisson Labs, Smithfield, US, USA).
5. Murashige and Skoog (MS) medium (MSP02, Caisson Labs, Smithfield, US, USA).
6. Petri plates.
7. 200 ml plastic containers with cap.

2.2 Seed Sterilization

1. Plant Preservative Mixture (PPM) (Plant Cell Technology, Washington, DC, USA).
2. 10% Sodium Hypochlorite (NaOCl) solution.
3. Forceps.
4. Screening caps.
5. 50 ml centrifuge tubes.

2.3 Plant Growth Media and Solutions (as Given in Bragg et al. [13]; Vogel and Hill [14])

1. Callus initiation media 1 (CIM1).

For 1 l:

Linsmaier and Skoog basal medium	4.43 g
Sucrose	30 g
Copper sulfate ($CuSO_4$) stock solution (0.6 mg/ml)	1 ml
Plant Preservative Mixture (PPM)	0.4 ml
Phytagel	2 g
2,4-D (5 mg/ml)[a]	0.5 ml

Set the pH to 5.8 with 0.1 N KOH.
[a]Add filter sterilized 2,4-D after autoclaving.

2. Callus initiation media 2 (CIM2)

For 1 l:

Linsmaier and Skoog basal medium	4.43 g
Sucrose	30 g
Copper sulfate ($CuSO_4$) stock solution (0.6 mg/ml)	1 ml
Plant Preservative Mixture (PPM)	0.4 ml
Phytagel	2 g
2,4-D (5 mg/ml)[a]	0.5 ml
Hygromycin B (50 mg/ml)[a]	1 ml

Set the pH to 5.8 with 0.1 N KOH.
[a]Add filter sterilized 2,4-D and hygromycin B after autoclaving

3. Regeneration media

For 1 l:

Linsmaier and Skoog basal medium	4.43 g
Maltose	30 g
Phytagel	2 g
Kinetin (0.2 mg/ml)[a]	1 ml

Set the pH to 5.8 with 0.1 N KOH.
[a]Add filter sterilized Kinetin after autoclaving.

4. Murashige and Skoog (MS) media

For 1 l:

Murashige and Skoog salts	4.42 g
Sucrose	30 g
Phytagel	2 g

Set the pH to 5.8 with 0.1 N KOH.

3 Method

This protocol is developed by modifying the original protocols by Bragg et al. [13] and Vogel and Hill [14] for large-scale regeneration and transposition of *Tnt1* retrotransposon in *Brachypodium*.

3.1 Seed Sterilization

1. Grow *Brachypodium* mother plant line 7-1 in 4″ pots containing synthetic soil mixture (Metro-Mix 830) in the greenhouse with daytime and nighttime temperatures of 22 °C and 18 °C, respectively. This line contains *Tnt1* retrotransposon in the background of Bd21-3 (Fig. 1). *Brachypodium* plants should be ready in 35–45 days to harvest mature caryopsis at seed filling stage (Fig. 2a, b).
2. Individual caryopsis can be harvested in 15 ml falcon tubes containing water.
3. Remove the lemma and palea with the help of forceps or manually from individual seeds and transfer those to a 50 ml falcon tube containing 4% solution of PPM. Once all the seeds are transferred to PPM solution shake the tube for 5–10 times and carefully rinse off the PPM solution using the screening caps without losing the seeds (*see* **Note 1**).
4. After rinsing off the PPM solution, add 10% bleach solution. Gently shake the tubes for 4 min on rotary/rocking shaker or manually.
5. Discard the bleach solution and rinse the seeds four times with sterile distilled water. At this point, seeds are surface sterilized and are ready for transfer to a sterile growth medium.

Fig. 1 The *Tnt1* cassette in pCAMBIA 1381xc binary vector which was used for transforming *Brachypodium* line Bd21-3 to generate transgenic mother plant used in this study

3.2 Callus Induction and Shoot Regeneration

1. Sterile seeds should be transferred to the Petri plates containing CIM1 media without any antibiotics. Make sure that the embryo side of the seeds is down touching the medium (Fig. 2c) (*see* **Note 2**).
2. The Petri plates are kept in an incubator set at 28 °C under dark conditions for 1 week.
3. After 1 week, seeds should be transferred to CIM2. About 3–4 weeks later, calli will emerge from the embryos and will be visible (Fig. 2d) (*see* **Note 3**). After 3 weeks, calli can be excised from the seeds and transferred to Petri plates containing fresh CIM media.
4. Calli should be grown for another 6 weeks on CIM media (Fig. 2e) by transferring them every 3 weeks to fresh CIM. By

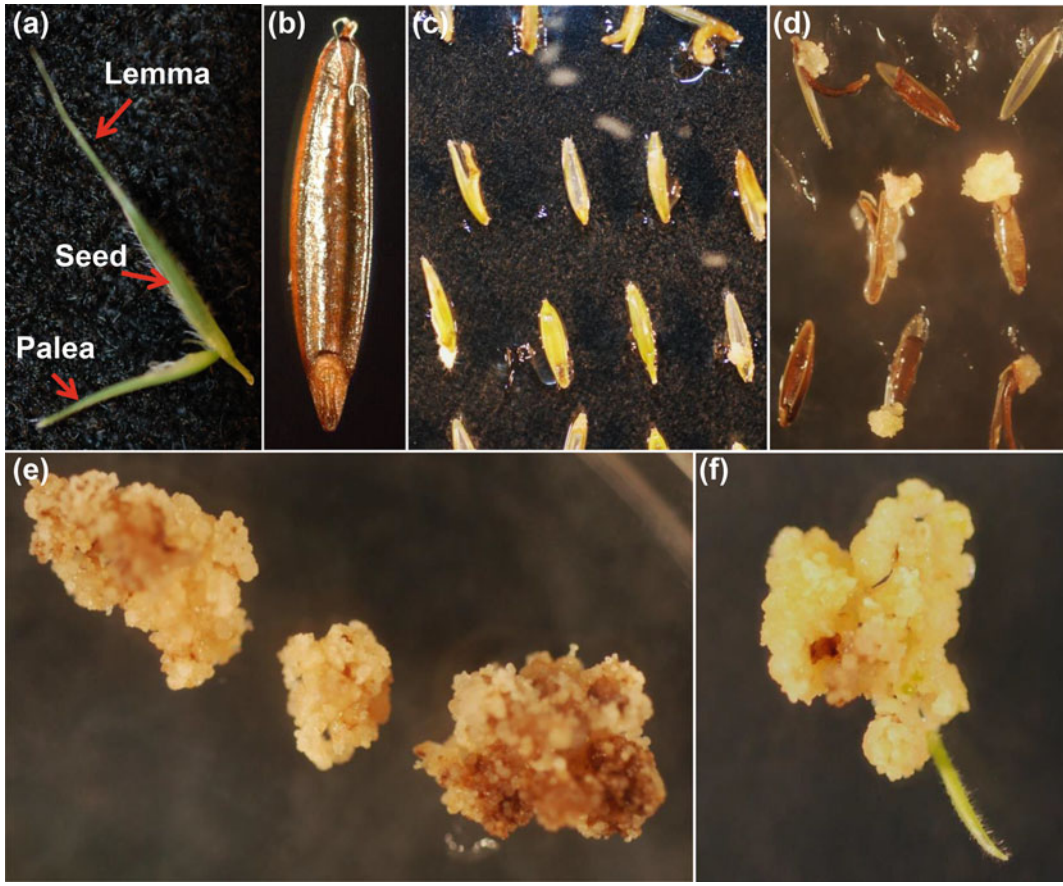

Fig. 2 Schematic representation of steps involved in large-scale regeneration of *Brachypodium Tnt1* insertion lines. (**a**) Harvested mature seed from 45 days old *Brachypodium* mother plant showing lemma and palea. (**b**) Seed after removal of lemma and palea. (**c**) Seeds placed on CIM media with the embryo side touching the media. (**d**) Callus regeneration 3–4 weeks after the initial culture. (**e**) Calli were carefully excised from seeds and placed on CIM media for another 6 weeks. (**f**) Regenerated shoot from the embryogenic callus

the end of 5th week, calli will be ready for transfer to shoot regeneration media.

5. Transfer the calli to Petri plates containing shoot regeneration medium.
6. At this point, Petri plates are kept in light (16/8 h day/night cycle) at 28 °C.
7. Once the shoots start to appear (Fig. 2f), transfer the plantlets to MS media in sterile 200 ml plastic containers until the roots are formed.
8. Transfer the plants into 4″ pots containing synthetic soil mixture (Metro-Mix 830). Put the pots in trays and cover them with dome to maintain high humidity. Place the plants in growth chamber or green house with 20 h/4 h day/night temperature of 22 °C/18 °C, respectively. These plants are R_0 plants and should contain multiple random insertions of *Tnt1* retrotransposon due to transposition during embryogenesis.
9. Isolate leaf tissues for molecular analysis and collect R_0 seeds from mature plants.

4 Notes

1. Treatment with PPM significantly reduces the fungal and other microbial contamination in the follow-up steps. We generally do not time this step because we did not notice any deleterious effect if the seeds are kept longer in PPM solution.
2. We found that the whole seeds produce good embryogenic calli and it eliminates the need of excising embryos from individual seeds.
3. One callus per embryo is considered as a unique event. Therefore, a single plant per callus per embryo should be maintained to avoid duplication of same transposition event.

Acknowledgement

We thank Janie Gallaway and Colleen Elles for taking care of the plants in the greenhouse. This work was supported by the Noble Research Institute, LLC.

References

1. Brkljacic J, Grotewold E, Scholl R et al (2011) *Brachypodium* as a model for the grasses: today and the future. Plant Physiol 157:3–13
2. Brutnell TP, Bennetzen JL, Vogel JP (2015) *Brachypodium distachyon* and *Setaria viridis*: model genetic systems for the grasses. Ann Rev Plant Biol 66:465–485

3. The International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768
4. Bragg JN, Wu J, Gordon SP, Guttman MA, Thilmony RL, Lazo GR, Gu YQ, Vogel JP (2012) Generation and characterization of the Western Regional Research Center *Brachypodium* T-DNA insertional mutant collection. PLoS One 7:e41916
5. Vain P, Worland B, Thole V, McKenzie N, Opanowicz M, Fish LJ, Bevan MW, Snape JW (2008) *Agrobacterium*-mediated transformation of the temperate grass *Brachypodium distachyon* (genotype Bd21) for T-DNA insertional mutagenesis. Plant Biotechnol J 6:236–245
6. Courtial B, Feuerbach F, Eberhard S, Rohmer L, Chiapello H, Camilleri C, Lucas H (2001) *Tnt1* transposition events are induced by *in vitro* transformation of *Arabidopsis thaliana*, and transposed copies integrate into genes. Mol Genet Genomics 265:32–42
7. Cui Y, Barampuram S, Stacey MG, Hancock CN, Findley S, Mathieu M, Zhang ZY, Parrott WA, Stacey G (2013) *Tnt1* retrotransposon mutagenesis: a tool for soybean functional genomics. Plant Physiol 161:36–47
8. d'Erfurth I, Cosson V, Eschstruth A, Lucas H, Kondorosi A, Ratet P (2003) Efficient transposition of the *Tnt1* tobacco retrotransposon in the model legume *Medicago truncatula*. Plant J 34:95–106
9. Duangpan S, Zhang W, Wu Y, Jansky SH, Jiang J (2013) Insertional mutagenesis using *Tnt1* retrotransposon in potato. Plant Physiol 163:21–29
10. Mazier M, Botton E, Flamain F, Bouchet JP, Courtial B, Chupeau MC, Chupeau Y, Maisonneuve B, Lucas H (2007) Successful gene tagging in lettuce using the *Tnt1* retrotransposon from tobacco. Plant Physiol 144:18–31
11. Tadege M, Wen JQ, He J, Tu HD, Kwak Y, Eschstruth A, Cayrel A, Endre G, Zhao PX, Chabaud M et al (2008) Large-scale insertional mutagenesis using the *Tnt1* retrotransposon in the model legume *Medicago truncatula*. Plant J 54:335–347
12. Grandbastien MA, Lucus H, Morel JB, Mhiri C, Vernhettes S, Casacubert JM (1997) The expression of the tobacco *Tnt1* retrotransposon is linked to the plant defense responses. Genetica 100:241–252
13. Bragg JN, Anderton A, Nieu R, Vogel JP (2015) Brachypodium distachyon. In: Wang K (ed) Agrobacterium protocols, Methods in molecular biology, vol 1223. Springer, New York, pp 17–33
14. Vogel JP, Hill T (2008) High-efficiency *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21-3. Plant Cell Rep 27:471–478

Chapter 6

Methods for Xyloglucan Structure Analysis in *Brachypodium distachyon*

Lifeng Liu

Abstract

Matrix-assisted laser desorption-ionization time-of-flight mass spectrometry (MALDI-TOF MS) has become an important tool for the analysis of biomolecules, such as DNA, peptides, and oligosaccharides. This technique has been developed as a rapid, sensitive, and accurate means for analyzing cell wall polysaccharide structures. Here, we describe a method using mass spectrometry to provide xyloglucan composition and structure information of Brachypodium plants which will be useful for functional characterization of xyloglucan biosynthesis pathway in *Brachypodium distachyon*.

Key words Mass spectrometry, Xyloglucan, Xyloglucan-endoglucanase, Oligosaccharides

1 Introduction

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is a soft ionization technique that causes minimal fragmentation of analytes compared with other ionization methods [1, 2]. This technique allows us to easily identify and analyze molecular ions even from a mixture and it has been widely used for analyzing biomolecules, such as DNA, peptides, and oligosaccharides [3–5].

MALDI-TOF MS has been successfully used in analyzing plant cell wall polysaccharide composition and structure, especially on xyloglucan which is a major hemicellulose in most dicot plants [6–10]. Specific enzymes, usually endoglucanases, were used to digest xyloglucan polysaccharides. Oligosaccharides released from polysaccharides are composed of several abundant xyloglucan structures [8]. Thus, mass spectra could be detected and used for analyzing the composition and structure of oligosaccharides which could represent the composition and structure of xyloglucan polysaccharides. This method could also be used to identify the decoration of xyloglucan, such as acetylation [11, 12]. MALDI-TOF MS is very sensitive and high throughput since only small amount of

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_6,

samples and very short run time (usually 1–2 min per sample) are required. This technique has been modified for xyloglucan mutant screening in Arabidopsis [13]. However, structure isomer cannot be distinguished by this technique. Other techniques, such as NMR and Dionex, need to be combined for detailed xyloglucan structure analysis [11, 12].

2 Materials

2.1 Plant Cell Wall Preparation

1. Murashige and Skoog (Sigma-Aldrich, cat. no. M5524-1L).
2. Sucrose (Sigma-Aldrich, cat. no. S0389-1KG).
3. Microcentrifuge tubes (Denville, cat. no. C2170) (*see* **Note 1**).
4. Stainless grinding ball (Retsch, cat. no. 22.455.0011) (*see* **Note 2**).
5. Ethanol (Fisher Scientific, cat. no. AC615090040).
6. Chloroform (Fisher Scientific, cat. no. C607SK-4).
7. Acetone (Fisher Scientific, cat. no. A18P-4).
8. Methanol (Fisher Scientific, cat. no. A454SK-4).

2.2 Enzyme Digest

1. Ammonium formate (Sigma-Aldrich, cat. no. 70221-25G-F).
2. Xyloglucan-specific endoglucanase (XEG, EC 3.2.1.151, 1 Unit of XEG releases 1 mmol xyloglucan oligosaccharides per min [14]) (*see* **Note 3**).

2.3 MALDI-TOF MS Analysis

1. 2,5-Dihydroxybenzoic acid (DHB, Sigma-Aldrich, cat. no. 85707-10MG-F).
2. Bio-Rad MSZ-501 (D) cation exchange resin beads (Bio-Rad, cat. no. 1427425) (*see* **Note 4**).
3. MALDI plate DE1580TA (Shimadzu, cat. no. TO-454R00).

3 Methods

3.1 Cell Wall Preparation

1. Place *Brachypodium distachyon* seeds (Bd21-3) on sterilized ½ Murashige and Skoog plates with 1% sucrose. Seal the plates with micropore tape and cover the plates with aluminum foil. Put the plates in cold room or 4 °C refrigerator for 7 days.
2. Transfer the covered plates into Brachypodium growth chamber (24 °C for 20 h and 18 °C for 4 h) for 3 days.
3. Harvest the shoot of germinated seeds (*see* **Note 5**). Place eight shoots and two grinding balls into 1.7 mL microcentrifuge tube and snap frozen in liquid nitrogen (*see* **Note 6**).

4. Grind the frozen material with a ball mixer mill (Retsch, MM400) at 25 Hz for 2 min.
5. Add 1 mL 70% ethanol into the tube and vortex briefly. Remove the balls with a magnet. Centrifuge the tube at 20,000 × *g* for 10 min in a tabletop centrifuge (Eppendorf 5417R). Remove the supernatant with vacuum or pipette.
6. Add 1 mL chloroform:methanol (1:1 v/v) and vortex the tubes. Make sure the materials are well suspended. Centrifuge the tube at 20,000 × *g* for 10 min and remove the supernatant.
7. Add 1 mL acetone and briefly vortex the samples. Centrifuge the tube at 20,000 × *g* for 5 min and remove the supernatant.
8. Dry the pellet in a concentrator (Eppendorf Vacufuge) for 30 min (*see* **Note 7**).

3.2 Enzyme Digestion of Xyloglucan

1. Add 1 mL 50 mM ammonium formate buffer (pH 4.5) and two grinding balls into the dried sample.
2. Grind the material with ball mixer mill at 25 Hz for 2 min and remove the balls with a magnet.
3. Incubate the tube at 250 rpm for 1 h in a 37 °C shaker (VWR).
4. Centrifuge the tubes and completely remove the supernatant with a vacuum or pipette.
5. Add 20 μL 50 mM ammonium formate buffer containing 0.2 U XEG to remaining pellet. Briefly vortex and incubate the tube at 250 rpm for 4 h in a 37 °C shaker (*see* **Note 8**).
6. Centrifuge the tube at 20,000 × *g* for 10 min in a tabletop centrifuge.
7. Transfer supernatant to a new 1.5 mL microcentrifuge tube and repeat **step 6** once.

3.3 MALDI-TOF MS Analysis

1. Transfer 10 μL digested supernatant into a new microcentrifuge tube.
2. Add approximately ten cation exchange beads to each sample and incubate at room temperature for 30 min.
3. Spot 2.5 μL DHB matrix (10 mg/mL in deionized water) onto MALDI plate and dry the matrix under vacuum with a pump (Welch2042 DryFast Ultra Pump) (*see* **Note 9**).
4. Spot 2.5 μL desalted sample liquid onto the dried DHB matrix spots (*see* **Note 10**).
5. Wait 2 min and dry the sample under vacuum.
6. Run the spotted samples with an MALDI-TOF mass spectrometer (Shimadzu AXIMA Performance). MALDI-TOF mass spectrometer is set with positive linear mode with accelerating

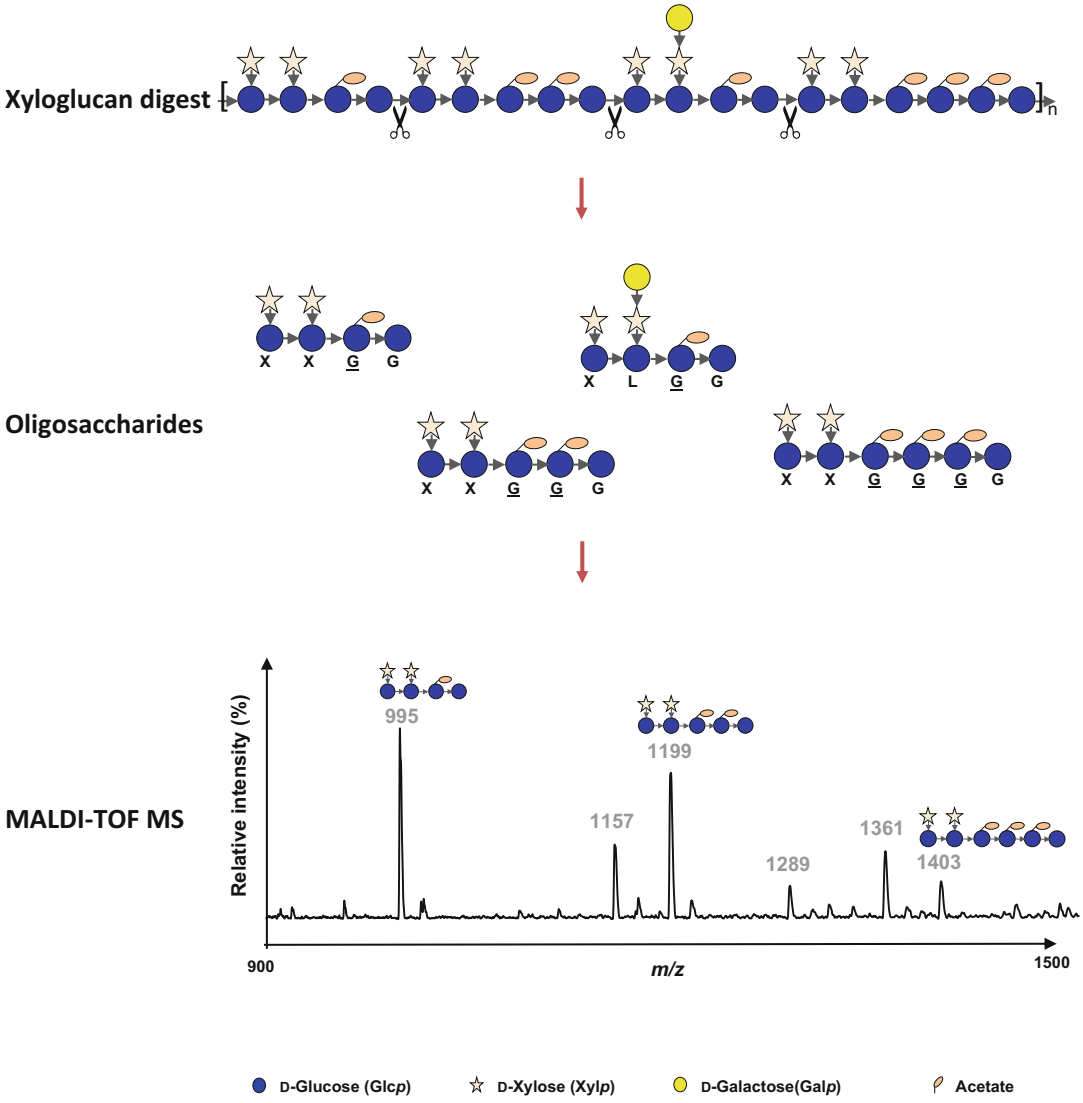

Fig. 1 Flow chart of xyloglucan analysis by MALDI-TOF MS. A representative xyloglucan oligosaccharides profile from the young shoot of *Bd21-3* plants was shown here. Detailed information of one letter label and *m*/*z* of xyloglucan oligosaccharides can be found in Table 1

voltage 20 kV. Selected mass range is 500–3500 Da with 200 spectra per sample (*see* **Note 11**).

7. A representative xyloglucan oligosaccharides profile is shown in Fig. 1 and detailed structure information can be found in Table 1 (*see* **Note 12**).

Table 1
Typical xyloglucan oligosaccharides detected in Brachypodium shoot

m/z	Composition	Suggested structure
953	P_2H_4	XXGG
995	$P_2H_4Ac_1$	XXGG
1115	P_2H_5	XXGGG/XLGG
1157	$P_2H_5Ac_1$	XXGGG/XXGGG/XLGG
1199	$P_2H_5Ac_2$	XXGGG/XLGG
1247	P_3H_5	XXXGG/XXLG/XLXG
1289	$P_3H_5Ac_1$	XXXGG/XXLG
1277	P_2H_6	XXGGGG, XLGGG
1319	$P_2H_6Ac_1$	XXGGGG/XXGGGG/XXGGGG/XLGGG/XLGGG/XLGGG
1361	$P_2H_6Ac_2$	XXGGGG/XXGGGG/XXGGGG/XLGGG/XLGGG/XLGGG
1403	$P_2H_6Ac_3$	XXGGGG/XLGGG

P pentose, *H* hexose, *Ac* acetate, *G* a backbone glucosyl residue without substitution, *X* additional xylosyl residue on the *O*-6 position of G, *L* X side-chain with additional galactosyl residue, *Underline* means glycosyl residue acetylated. Detailed nomenclature is described by Schultink et al. [15] and Tuomivaara et al. [16]. *Subscript numbers in the second column* indicate the number of residues. Possible structural isomers are listed here

4 Notes

1. Eppendorf 2 mL tubes are an alternative option. However, 1.7 mL tubes are better for supernatant removal after centrifugation without disturbing the pellet. Tubes with volume less than 1.5 mL are not recommended.
2. Grinding with larger balls (diameter larger than 5 mm) is not recommended, as those balls will easily break the frozen tubes during grinding.
3. The XEG used here is a gift from Kirk Schorr (Novozymes). There are commercial xyloglucan endoglucanases available (e.g. Megazyme, cat. no. E-XEGP).
4. The MSZ 501 (D) resin purchased is a mixture of both anion and cation resin beads. Only cation beads were used here. To remove the anion beads, put 100 mL mixed beads and 300 mL water into a 500 mL beaker and swirl the beaker. Cation beads with golden brown color will gather at the bottom. Carefully pour out the top layer with blue beads and resuspend beads with 300 mL water. Repeat several times until you have pure cation beads. Store the cation beads in water and wash the beads with more than three times of water each time before use.

5. Other tissues could also be used but the amount of xyloglucan is too low to detect. Dark-grown young tissues work better based on our experience.
6. The sampling and freezing steps should be done as quickly as possible. You can also freeze-dry the collected samples prior to cell wall material preparation. The dried material could be stored at room temperature for several months.
7. You can also dry the acetone washed samples in a chemical hood at room temperature. But it takes longer time.
8. Longer digestion time, such as 8 h or overnight also works but sodium azide (NaN_3, 0.2% w/v) needs to be added to prevent bacterium growth.
9. A stock solution of DHB matrix can be stored at −20 °C freezer and dark condition for several months. Thoroughly thaw and vortex the matrix. Make sure it is completely dissolved before use.
10. We recommend using the same amount of sample and DHB. The volume of each spot should not exceed 4 μL. You can dry and re-dissolve the desalted sample in water with less volume (~3 μL) to concentrate released oligosaccharides for better signals.
11. A representative spectrum is generated with the average of total spectra. MALDI-TOF MS cannot be used for precise quantification. Relative abundance of oligosaccharides is calculated based on the ratio of the height of selected ion to all ions of interest.
12. DHB usually produces $[M + Na]^+$ ion. Ionized oligosaccharides will be analyzed as for their sodium adducts with an additional +23 m/z. However, DHB could also produce very weaker $[M + K]^+$ ion (an additional +39 m/z). If oligo-ladder was found in mass spectra, lichenase (EC 3.2.1.73) needs to be added to remove potential mix-linked glucan contamination before XEG digest.

Acknowledgement

This work was funded by the Energy Bioscience Institute, University of California, Berkeley. The author would like to thank Kirk Schorr (Novozymes) for providing the enzymes XEG.

References

1. Karas M, Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 60(20):2299–2301
2. Calderaro A, Arcangeletti MC, Rodighiero I, Buttrini M, Gorrini C, Motta F, Germini D, Medici MC, Chezzi C, De Conto F (2014) Matrix-assisted laser desorption/ionization

time-of-flight (MALDI-TOF) mass spectrometry applied to virus identification. Sci Rep 4:6803. doi:10.1038/srep06803

3. Tang K, Taranenko NI, Allman SL, Cháng LY, Chen CH (1994) Detection of 500-nucleotide DNA by laser desorption mass spectrometry. Rapid Commun Mass Spectrom 8 (9):727–730. doi:10.1002/rcm.1290080913
4. Wu KJ, Steding A, Becker CH (1993) Matrix-assisted laser desorption time-of-flight mass spectrometry of oligonucleotides using 3-hydroxypicolinic acid as an ultraviolet-sensitive matrix. Rapid Commun Mass Spectrom 7 (2):142–146. doi:10.1002/rcm.1290070206
5. Webster J, Oxley D (2012) Protein identification by MALDI-TOF mass spectrometry. Methods Mol Biol 800:227–240. doi:10.1007/978-1-61779-349-3_15
6. Bauer S (2012) Mass spectrometry for characterizing plant cell wall polysaccharides. Front Plant Sci 3:45. doi:10.3389/fpls.2012.00045
7. Günl M, Gille S, Pauly M (2010) OLIgo mass profiling (OLIMP) of extracellular polysaccharides. J Vis Exp (40). doi:10.3791/2046
8. Günl M, Kraemer F, Pauly M (2011) Oligosaccharide mass profiling (OLIMP) of cell wall polysaccharides by MALDI-TOF/MS. Methods Mol Biol 715:43–54. doi:10.1007/978-1-61779-008-9_3
9. Obel N, Erben V, Schwarz T, Kühnel S, Fodor A, Pauly M (2009) Microanalysis of plant cell wall polysaccharides. Mol Plant 2(5):922–932. doi:10.1093/mp/ssp046
10. Pauly M, Gille S, Liu L, Mansoori N, de Souza A, Schultink A, Xiong G (2013) Hemicellulose biosynthesis. Planta 238(4):627–642. doi:10.1007/s00425-013-1921-1
11. Gille S, de Souza A, Xiong G, Benz M, Cheng K, Schultink A, Reca IB, Pauly M (2011) O-acetylation of Arabidopsis hemicellulose xyloglucan requires AXY4 or AXY4L, proteins with a TBL and DUF231 domain. Plant Cell 23 (11):4041–4053. doi:10.1105/tpc.111.091728
12. Liu L, Paulitz J, Pauly M (2015) The presence of fucogalactoxyloglucan and its synthesis in rice indicates conserved functional importance in plants. Plant Physiol. doi:10.1104/pp.15.00441
13. Lerouxel O, Choo TS, Séveno M, Usadel B, Faye L, Lerouge P, Pauly M (2002) Rapid structural phenotyping of plant cell wall mutants by enzymatic oligosaccharide fingerprinting. Plant Physiol 130(4):1754–1763. doi:10.1104/pp.011965
14. Pauly M, Andersen LN, Kauppinen S, Kofod LV, York WS, Albersheim P, Darvill A (1999) A xyloglucan-specific endo-beta-1,4-glucanase from Aspergillus aculeatus: expression cloning in yeast, purification and characterization of the recombinant enzyme. Glycobiology 9 (1):93–100
15. Schultink A, Liu L, Zhu L, Pauly M (2014) Structural diversity and function of xyloglucan sidechain substituents. Plants 3(4):526–542
16. Tuomivaara ST, Yaoi K, O'Neill MA, York WS (2015) Generation and structural validation of a library of diverse xyloglucan-derived oligosaccharides, including an update on xyloglucan nomenclature. Carbohydr Res 402:56–66. doi:10.1016/j.carres.2014.06.031

Chapter 7

Genomic Approaches to Analyze Alternative Splicing, A Key Regulator of Transcriptome and Proteome Diversity in *Brachypodium distachyon*

Sonia Irigoyen, Renesh H. Bedre, Karen-Beth G. Scholthof, and Kranthi K. Mandadi

Abstract

Alternative splicing (AS) promotes transcriptome and proteome diversity in plants, which influences growth and development, and host responses to stress. Advancements in next-generation sequencing, bioinformatics, and computational biology tools have allowed biologists to investigate AS landscapes on a genome-wide scale in several plant species. Furthermore, the development of *Brachypodium distachyon* (Brachypodium) as a model system for grasses has facilitated comparative studies of AS within the *Poaceae*. These analyses revealed a plethora of genes in several biological processes that are alternatively spliced and identified conserved AS patterns among monocot and dicot plants. In this chapter, using a Brachypodium-virus pathosystem as a research template, we provide an overview of genomic and bioinformatic tools that can be used to investigate constitutive and alternative splicing in plants.

Key words Alternative splicing, Brachypodium, *Panicum mosaic virus*, RNA-sequencing, Bioinformatics

1 Introduction

Alternative splicing (AS) has important biological consequences in plant growth and development, flowering, circadian clock function, and stress responses [1, 2]. In recent years, RNA sequencing (RNA-seq) technology has emerged as an invaluable tool for transcriptome and AS analyses in multiple plant species. RNA-seq has several advantages over conventional approaches such as microarrays [3–5]. RNA-seq does not require prior sequence knowledge of the genome, genes, or transcripts. In addition to discovering new genes and splice-variants, RNA-seq can be used to validate and improve existing gene and genome annotations. RNA-seq technology provides much higher resolution for gene expression measurements, is sensitive in estimating expression of low abundance genes,

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_7,

and has a broad dynamic range to detect differentially expressed genes among samples. Importantly, RNA-seq allows estimation of isoform-level expression, which can be utilized to determine AS patterns on a genome-wide scale.

We have recently analyzed the transcriptome and AS patterns of *Brachypodium distachyon* (Brachypodium) infected with *Panicum mosaic virus* (PMV) and its satellite virus (SPMV) using the RNA-seq approach [6–8]. We identified ~44,000 transcripts using the Tuxedo RNA-seq data analysis pipeline [9] and found that approximately 42% of the intron-containing transcripts were alternatively spliced [7]. The major types of AS events in eukaryotes are classified as exon-skipping, intron-retention, alternate acceptor, and alternate donor types [1, 2]. Complex AS events can occur, to include duplications or combinations of the aforementioned four types. Exon skipping events are predominant (> 40%) in animals, whereas intron-retention events predominate (~40%) in plants [10]. Of the ~9541 AS events detected in Brachypodium, approximately 36%, 27%, 14%, 9%, and 14% were intron-retention, alternate acceptor, alternate donor, exon skipping, and complex events, respectively [7]. The Brachypodium AS ratios are comparable both to other grasses such as rice, maize and sorghum, and to dicot plants such as Arabidopsis, potato, Medicago, and poplar [7]. PMV and PMV+SPMV infections did not significantly alter the overall ratios of AS types; however, higher numbers of AS events in different categories were observed as a consequence of virus infection [7].

Multiple genes encoding protein kinases, resistance proteins, transcription, and splicing factors were differentially spliced during PMV and PMV+SPMV infection [7]. We further characterized a splicing factor, Bd-*SCL33*, whose splicing patterns are conserved with a distant ortholog in Arabidopsis [7, 8, 11]. Using reverse transcription polymerase chain reaction (RT-PCR) analysis, we validated the presence of at least six splice variants for Bd-*SCL33*, and further demonstrated that they were developmentally regulated [11]. Several Bd-*SCL33* splice variants, containing segments of intron I and III, were mis-expressed in Brachypodium plants challenged with seven different viruses, including PMV [7, 8]. Cloning and sequencing of the six Bd-*SCL33* splice variants revealed multiple premature stop codons in their open reading frames, resulting in truncated proteins lacking portions of the N-terminal RNA recognition motif (RRM)-domain and/or a Ser/Arg-rich (SR) region [7, 8]. These putative truncated Bd-SCL33 proteins could potentially compete with wild-type Bd-SCL33 in a dominant negative manner and interfere with its native function. In turn, this could affect AS landscapes mediated by Bd-SCL33 during stress and development [11].

In summary, we found extensive changes in genome-wide AS landscapes, which are predicted to affect transcriptome and proteome diversity of Brachypodium during biotic stress. These data underscore the utility of RNA-seq approaches to limn AS patterns

in plants. We used several bioinformatics programs and computational tools, many of which are open-source, to perform RNA-seq data analysis. Additionally, community cyberinfrastructure resources such as iPlant and Galaxy make it possible for individuals to perform large-scale data analysis using cloud-based computing [12, 13]. The aim of this chapter is to outline a typical workflow for AS analysis using open-source bioinformatics programs.

2 Materials

2.1 System Requirements

64-bit (Linux/Unix or Mac OS X) computer.

At least 30 GB (or more depending on the RNA-seq data size) of free disk space.

Approximately 8 GB (16 GB or more are preferred) of RAM.

2.2 Public RNA-seq Repositories and Reference Genomes

NCBI-SRA (http://www.ncbi.nlm.nih.gov/sra) (Last accessed date: Feb 1, 2017).

European Nucleotide Archive (http://www.ebi.ac.uk/ena) (Last accessed date: Feb 1, 2017).

Reference genomes and annotations (http://plants.ensembl.org/index.html and https://phytozome.jgi.doe.gov/pz/portal.html) (Last accessed date: Feb 1, 2017).

2.3 RNA-seq Quality Control and Data Analysis Tools

FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Last accessed date: Feb 1, 2017).

NGS QC Toolkit (http://www.nipgr.res.in/ngsqctoolkit.html) (Last accessed date: Feb 1, 2017).

Sabre (https://github.com/najoshi/sabre) (Last accessed date: Feb 1, 2017).

Scythe (https://github.com/vsbuffalo/scythe) (Last accessed date: Feb 1, 2017).

FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) (Last accessed date: Feb 1, 2017).

Bowtie (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) (Last accessed date: Feb 1, 2017).

TopHat (https://ccb.jhu.edu/software/tophat/index.shtml) (Last accessed date: Feb 1, 2017).

Cufflinks (http://cole-trapnell-lab.github.io/cufflinks/) (Last accessed date: Feb 1, 2017).

StringTie (https://ccb.jhu.edu/software/stringtie/) (Last accessed date: Feb 1, 2017).

CummeRbund (http://compbio.mit.edu/cummeRbund/) (Last accessed date: Feb 1, 2017).

2.4 Alternative Splicing and Data Visualization Tools

AStalavista (http://genome.crg.es/astalavista/) (Last accessed date: Feb 1, 2017).

JuncBASE (http://compbio.berkeley.edu/proj/juncbase/Home.html) (Last accessed date: Feb 1, 2017).

MISO tools (http://genes.mit.edu/burgelab/miso/) (Last accessed date: Feb 1, 2017).

SeqMonk (http://www.bioinformatics.babraham.ac.uk/projects/seqmonk/) (Last accessed date: Feb 1, 2017).

Integrative Genomics Viewer (https://www.broadinstitute.org/igv/) (Last accessed date: Feb 1, 2017).

2.5 Community Cyberinfrastructure Resources

iPlant Discovery Environment (http://www.cyverse.org/discovery-environment) (Last accessed date: Feb 1, 2017).

iPlant Atmosphere (https://atmo.cyverse.org/application/images) (Last accessed date: Feb 1, 2017).

Galaxy (https://galaxyproject.org/) (Last accessed date: Feb 1, 2017).

GenomeSpace (http://www.genomespace.org/) (Last accessed date: Feb 1, 2017).

3 Methods

3.1 Experimental Design and Considerations

3.1.1 Technical and Biological Replicates

If all quality control steps are rigorously followed, RNA-seq produces little technical variability when compared to microarrays [9]. However, since biological variation persists, it is advisable to include multiple biological replicates in the experimental design. A minimum of three biological replicates per condition is preferred, however the exact number should be determined based on the extent of biological variation in each system. RNA-seq analysis of the biological replicate samples allows for an estimation of dispersion in gene expression among the replicates, which can be used to calculate "mathematically averaged" changes in gene expression. As an alternative, cost-effective approach, equal amounts of RNA from the biological replicates are pooled before subjecting the samples to RNA-seq in a "biologically-averaged" design [14, 15]. Comparative analysis of RNA-seq data of mathematically versus biologically averaged replicates showed that much of the differential gene expression changes can be similarly identified in both approaches [15]. However, if the cost of sequencing is not a major concern or by using multiplexing approach to reduce the cost for replicate sequencing, individual biological replicates and mathematical averaging is advisable, to fully utilize the potential of downstream bioinformatics programs and to estimate subtle, yet statistically significant, differences in gene and transcript abundance.

3.1.2 Sequencing Strategy

Among the RNA-seq strategies, there is the choice of single-end or paired-end sequencing. If the goal of the experiment is to measure expression changes at the gene level, single-end RNA-seq is sufficient. However, to determine transcript-level expression and for AS analysis, paired-end RNA sequencing is recommended [4, 7, 9]. Long (~150 bases), paired-end reads improve the transcript assembly process, and aid in the identification of novel transcripts, splice variants, and gene fusions. The depth of sequencing (typically 20–60 million reads per sample) can vary depending on the size of the genome and ploidy levels, and should be determined accordingly [16].

3.1.3 RNA-Quality and Library Preparation

High-quality RNA is a prerequisite for successful RNA-seq, as it impacts downstream bioinformatics analysis. Several commercial RNA extraction kits (e.g., Qiagen, Zymo Research) are available that use spin-column based procedures to selectively purify high-quality, inhibitor-free RNA suitable for RNA-seq. Many of the kits include an in-column DNase treatment to eliminate residual genomic DNA contamination. After isolation, the RNA quality and quantity should be empirically assessed by denaturing RNA gel electrophoresis or with a bioanalyzer. The preparation of high-quality RNA-seq libraries is crucial as it affects the quality of RNA-seq reads, and is one of the major sources of technical variability and systematic bias in downstream data analysis [9]. Depending on the RNA-seq platform, the library preparation should be stringently optimized following manufacturer's instructions.

3.2 RNA-seq Data Processing and AS Analysis

3.2.1 Preprocessing Raw RNA-seq Libraries

RNA-seq services obtained through commercial or academic core facilities often provide demultiplexed and preprocessed reads, devoid of adaptors, primers, and/or barcodes. If pre-processing is not included, programs such as Sabre and Scythe can be used to clean the libraries (Fig. 1, Table 1). It is also critical to determine the quality of the RNA-seq reads before proceeding to RNA-seq data analysis. Programs such as FastQC and NGS QC Toolkit [17] can be used to generate quality reports, and accordingly RNA-seq libraries have to be filtered and trimmed to eliminate poor quality reads and bases with low Phred scores. The FASTQ quality filter and FASTQ quality trimmer, components of the FASTX-toolkit, can be used to perform filtering and trimming, respectively.

3.2.2 Tuxedo RNA-seq Analyses Pipeline

For organisms with an available reference genome, the Tuxedo pipeline is highly suitable to perform simultaneous RNA-seq mapping to reference sequences, transcript assembly, differential expression at gene and isoform level, and splicing analyses [9].

(a) *Map reads to genome using TopHat.* Quality filtered reads of each sample are separately aligned to a reference genome using TopHat (Fig. 1, Table 1) [5, 9]. TopHat overcomes

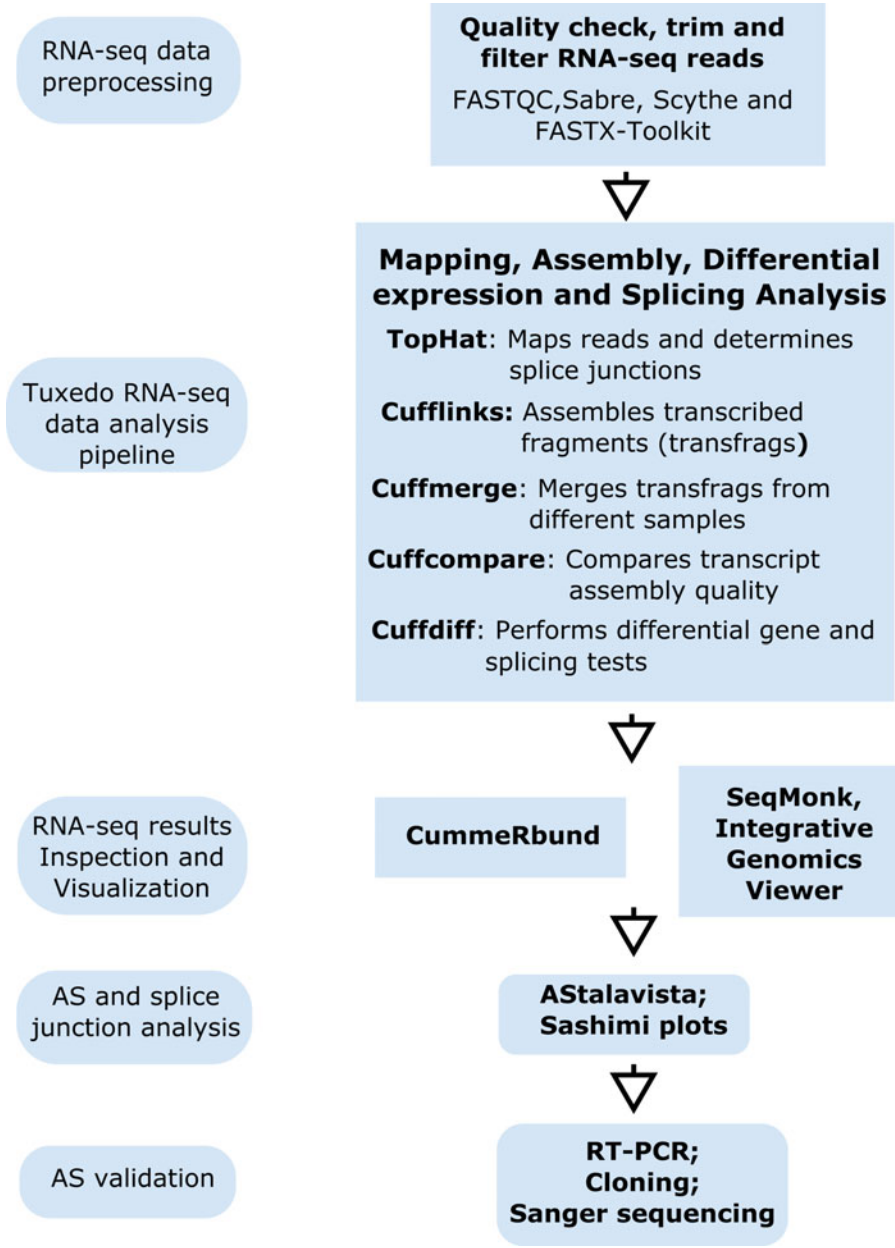

Fig. 1 Workflow of RNA-sequencing and alternative splicing (AS) analyses

the limitation of Bowtie to align the reads over the splice junctions. Briefly, this program uses the ultra-high throughput short read aligner, Bowtie [18] to align the RNA-seq reads to transcriptome and genome. The segmental mapping (splitting the reads into smaller segments and mapping to genome) of transcriptomic and genomic unmapped reads is

Table 1
Input and output file types for the RNA-seq and AS programs

Program	Input files	Output files
Sabre	.fastq (raw reads)	.fastq
Scythe	.fastq (raw reads)	.fastq
FASTQ quality filter	.fastq (raw reads)	.fastq
FASTQ quality trimmer	.fastq (raw reads)	.fastq
Tophat	.fastq (processed reads) .fasta (reference genome) .gtf (reference annotation)	.bam (alignment file) .bam.bai (alignment index file) .bed (junction file)
Cufflinks	.bam (alignment file) .gtf (reference annotation)	.gtf (transcript assembly)
Cuffmerge	.gtf (transcript assembly) .gtf (reference annotation)	.gtf (merged transcript annotations)
Cuffdiff	.bam (read alignments) .gtf (merged transcript annotations)	.diff (differential promoter, splicing and expression results)
Integrated Genomics Viewer, Seqmonk, CummeRbund	.fasta (reference genome) .gtf (merged transcript annotations) .bam (read alignments) .bed (splice junctions) Cuffdiff output (for CummeRbund)	Visualization and image export
AStalavista	.gtf (merged or sample specific transcript annotations) .fasta (reference genome)	.gtf (AS landscapes)
Sashimi plots	.fasta (reference genome) .gtf (merged transcript annotations) .bam (read alignments) .bed (splice junctions)	Visualization and image export

used to successfully determine potential splice junctions [19]. The later feature of TopHat makes it more suitable for discovery of novel AS transcripts and gene fusions.

(b) *Assembling transcripts using Cufflinks.* The TopHat produced alignments are then used to assemble individual transcribed fragment (transfrags) from RNA-seq reads using Cufflinks (Fig. 1, Table 1) [9]. It is often challenging to determine all splice variants of a gene within a sample owing to multiple isoforms for a given gene and difficulty in estimating the origin of the reads from different splice variants. To address this, the parsimonious algorithm of Cufflinks generates a minimum number of transcripts from all reads in each splice graph, i.e., it determines as few full-length transfrags that are needed to "explain" all the splicing events

in the mapped data [9]. After the initial transfrag assembly, Cufflinks quantifies normalized expression values (fragments per kilobase of transcript per million mapped fragments [FPKM]) of transfrags, and filters low abundance artifactual transfrags such as immature pre-mRNAs that are very low in abundance as compared to the matured transcripts [20].

In addition to Cufflinks, the StringTie transcript assembler can be used to derive isoforms of a gene using the TopHat produced spliced read alignment files [21]. StringTie integrates both genome-guided and *de novo* assembly approaches to assemble the transcripts. When compared to Cufflinks, StringTie is a faster algorithm that assembles and quantifies expression levels of transcripts simultaneously [21]. The output obtained by StringTie can be further used for differential expression analysis by Cuffdiff [20].

(c) *Merge and compare transcript assemblies using Cuffmerge and Cuffcompare:* Before differential expression analysis, it is critical to obtain a comprehensive set of transfrags present among all replicates and samples. This is necessary to have a uniform basis for calculating gene and transcript expression in each sample [9, 20]. To avoid the pooling of the sequencing reads from RNA-seq samples and replicates, which can be computationally expensive and increases the probability to assemble transcripts incorrectly, the Tuxedo pipeline provides a program Cuffmerge which merges the assembled transfrags from Cufflinks from different RNA-seq samples parsimoniously. Cuffmerge can be used to merge individual transfrag assemblies generated from Cufflinks, and can also include reference transcript into the final merged assembly, if a reference genome annotation is provided (Fig. 1, Table 1). Cuffmerge also aids in the recovery of complete gene sequences for low abundant genes that may have poor sequencing coverage in individual replicates or samples. The merged annotation file obtained from Cuffmerge is used for screening differentially expressed genes and transcripts using Cuffdiff [9].

After merging transcript assemblies, a quality check is recommended to assess the transcript assembly quality, particularly by comparing to the reference transcript annotations. Furthermore, the low sequencing depth usually produces partial transcripts from Cufflinks assembly program and it is challenging to differentiate full length versus partially reconstructed transcripts. To identify novel transcripts from known transcripts, the Cufflinks suit includes a program called Cuffcompare (compare Cufflinks produced assemblies with reference annotation) which can run on individual samples as well as merged annotation files [9].

Cuffcompare allows determination of base-, exon-, intron-, transcript-, and locus-level sensitivity and specificity for the transcript assemblies [9].

(d) *Differential expression and splicing analysis with Cuffdiff.* Although several count-based programs (e.g., DESeq, edgeR) are available to perform differential gene expression analysis, Cuffdiff is particularly well-suited to conduct simultaneous analyses such as differential gene and transcript expressions, as well as differential splicing and promoter analyses [20]. Cuffdiff determines differentially spliced genes by parsing the transcripts into groups based on the transcriptional start site (TSS). In addition, Cuffdiff controls the two key problems in RNA-seq experiments, (a) uncertainty in read count (ambiguously mapped sequences reads to different transcripts) and (b) count overdispersion (variability in sequence read counts across sample replicates), by using a beta and negative binomial distributions [20]. Thus, Cuffdiff allows for the identification of statistically significant changes in expression both at the transcriptional and post-transcriptional levels. Cuffdiff produces tabular output files for differential expression analysis which can be viewed and sorted for statistically significant differences using spreadsheet files (Fig. 1, Table 1).

3.2.3 Global Visualization, Inspection, and Plotting of Transcriptome and AS Data

(a) *CummeRbund.* Global visualization and inspection of transcriptome results can be performed using the CummeRbund program [9, 22]. This package is an interphase made to simplify exploration (plotting and clustering) of gene and transcript expression data produced by Cuffdiff (Fig. 1, Table 1). Expression density plots, box plots, scatter plots, and dispersion plots can be constructed using CummeRbund, to determine the extent of systematic bias and to assess the quality of the transcriptome analysis. Significant differences in gene and transcript abundance can be plotted on a genome-wide scale using volcano plots. Further, multivariate clustering and statistical relationships can be evaluated using principal component and distance-matrix plots. The various results of CummeRbund analyses can be exported as publication-ready images.

(b) *SeqMonk and Integrative Genomics Viewer.* Chromosome-level visualization and analysis of RNA-seq reads in the context of the reference genome can be performed using SeqMonk (Babraham Bioinformatics) and Integrative Genomics Viewer [23]. The RNA-seq binary format alignment files (.bam), reference genome (.fasta), and genome annotations (.gtf files) are used to perform quantitation of the read densities (RPM or RPKM), and to generate chromosome-level

plots (Fig. 1, Table 1). The chromosome visualization tools are also useful to query transcript junctions, as well as to extract selected sequences.

3.2.4 Alternative Splicing Landscape and Splice Junction Analysis Tools

(a) *Alternative splicing transcriptional landscape visualization tool* (AStalavista). The extent of the different types of AS occurring on a genome-wide scale can be determined using AStalavista [24, 25]. The Cufflinks transcript annotations and splice junction coordinates (.gtf) are used as input for AStalavista (Fig. 1, Table 1). Using the provided exon–intron junctions of all transcripts, AStalavista dynamically extracts and classifies various AS events as intron-retention, exon-skipping, alternate acceptor and alternate donor, and other complex types. AStalavista can output these alternative splicing distributions as pie charts and grouped lists based on the exon–intron structures (.gtf). Alternatively, JuncBase, a junction-based splicing analysis tools can be used to characterize alternative splicing events from the RNA-seq data [26].

(b) *Sashimi plots.* Sashimi plot, a part of the MISO (Mixture of ISOforms) framework, is a useful tool for visualization and analyses of splice junctions of AS transcripts [27]. Multiple samples can be compared simultaneously and read support can be quantitated by plotting RNA-seq read densities across the splice junctions (Fig. 1, Table 1). According to a user-defined threshold for read coverage, the various junctions among AS transcripts can be determined and visually represented as loops.

3.2.5 Working Example to Identify Differentially Expressed Genes and Alternative Transcript Events

Below are typical commands for RNA-seq and alternative splicing analysis that can be customized for optional parameters as per user requirements:

(a) *Tophat: Map cleaned RNA-seq reads to reference genome sequence.*

```
$ tophat -p 10 -G ref_genes.gtf -o sample1_map_out
ref_genome_bowtie_index sample1_R1.fastq
sample1_R2.fastq
$ tophat -p 10 -G ref_genes.gtf -o sample2_map_out
ref_genome_bowtie_index sample2_R1.fastq
sample2_R2.fastq
```

(b) *Cufflinks: Assemble the mapped RNA-seq reads to genes and transcripts.*

```
$ cufflinks -p 10 -o sample1_cufflink_out
sample1_map_out/accepted_hits.bam
$ cufflinks -p 10 -o sample2_cufflink_out
sample2_map_out/accepted_hits.bam
```

(c) *Cuffmerge: Create a merged transcriptome from the cufflinks assembled reads.*

```
cuffmerge -p 10 -g ref_genes.gtf -s ref_genome.fasta
-o sample_cuffmerge_out assemblies.txt
```

(Note: Assemblies.txt file created by adding path of cufflink assembled transcripts.gtf file each line for each sample)

(d) *Cuffdiff: Perform differential gene and transcript expression analysis.*

```
$ cuffdiff -o cuffdiff_out -b ref_genome.fasta -p 10
-L S1,S2 -u sample_cuffmerge_out/merged.gtf
sample1_map_out/accepted_hits.bam
sample2_map_out/accepted_hits.bam
```

(Note: differentially expressed genes and transcripts obtained from cuffdiff can be further visualized using CummeRbund tool)

(e) *Astalavista: Identify alternative splicing events.*

```
$ astalavista -t asta -i
sample1_cufflink_out/transcript.gtf -d 0
$ astalavista -t asta -i
sample2_cufflink_out/transcript.gtf -d 0
```

3.3 Validation of AS Patterns

3.3.1 RT-PCR, Cloning and Sequencing

Validation of AS results by wet-lab approaches is critical before pursuing further studies. To this end, selected AS events and splice variants can be analyzed by reverse-transcription polymerase chain reaction (RT-PCR), followed by cloning and Sanger-based DNA sequencing of the different AS variants to determine the sequence of the splice variants, and of the encoded proteins (Fig. 1, Table 1). Lastly, the biological significance of the AS events is determined by hypothesis-driven, reverse genetic approaches that are beyond the scope of this chapter.

4 Notes

The cost of RNA-seq has decreased considerably in recent years, largely due to improvements in sequencing chemistries and multiplexing capacities of the sequencing instruments. However, analyzing large-scale next-generation sequencing (NGS) data sets is a challenge for wet-lab researchers, given that data analysis is computationally intensive, requiring dedicated workstations and high-performance computers (HPC). Moreover, several analyses require researchers to possess considerable knowledge in bioinformatics and command-line interfaces. Cyberinfrastructure initiatives such as CyVerse (formerly iPlant Collaborative) and Galaxy are

addressing these challenges by making cloud-based computing resources and bioinformatics tools accessible with user-friendly interfaces [12, 13]. Several bioinformatic programs described here are accessible through CyVerse and Galaxy, and we recommend these open-source cyberinfrastructure resources to researchers who are keen to query their NGS data. As with any bioinformatics programs, we recommend users to be aware of updated versions and new software, study the program manuals, select parameters and commands that are appropriate for the specific experiment, and perform the recommended quality controls of the data. With a thoughtful strategy and selection of appropriate RNA-seq data analysis resources, biologists are today well positioned to ascertain plant AS phenomenon at an unprecedented level.

Acknowledgment

This study was supported by funds from USDA-NIFA-AFRI (2016-67013-24738) to K-B.G.S. and K.K.M., Texas A&M AgriLife Research Bioenergy/Bioproducts Grant (124738-96210) to K.K.M.

References

1. Staiger D, Brown JWS (2013) Alternative splicing at the intersection of biological timing, development, and stress responses. Plant Cell 25(10):3640–3656. doi:10.1105/tpc.113.113803
2. Reddy ASN, Marquez Y, Kalyna M, Barta A (2013) Complexity of the alternative splicing landscape in plants. Plant Cell 25 (10):3657–3683. doi:10.1105/tpc.113.117523
3. Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, Guernec G, Jagla B, Jouneau L, Laloë D, Le Gall C, Schaëffer B, Le Crom S, Guedj M, Jaffrézic F (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14(6):671–683. doi:10.1093/bib/bbs046
4. Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7(12):1009–1015. doi:10.1038/nmeth.1528
5. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111. doi:10.1093/bioinformatics/btp120
6. Mandadi KK, Scholthof K-BG (2015) Genomic architecture and functional relationships of intronless, constitutively- and alternatively-spliced genes in *Brachypodium distachyon*. Plant Signal Behav 10(8):e1042640. doi:10.1080/15592324.2015.1042640
7. Mandadi KK, Scholthof K-BG (2015) Genome-wide analysis of alternative splicing landscapes modulated during plant-virus interactions in *Brachypodium distachyon*. Plant Cell 27:71–85. doi:10.1105/tpc.114.133991
8. Mandadi KK, Pyle JD, Scholthof K-BG (2015) Characterization of SCL33 splicing patterns during diverse virus infections in *Brachypodium distachyon*. Plant Signal Behav 10(8):e1042641. doi:10.1080/15592324.2015.1042641
9. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578. doi:10.1038/nprot.2012.016
10. Liu R, Loraine AE, Dickerson JA (2014) Comparisons of computational methods for

differential alternative splicing detection using RNA-seq in plant systems. BMC Bioinformatics 15(1):364

11. Thomas J, Palusa SG, Prasad KVSK, Ali GS, Surabhi G-K, Ben-Hur A, Abdel-Ghany SE, Reddy ASN (2012) Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. Plant J 72(6):935–946. doi:10.1111/tpj.12004
12. Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, Muir A, Merchant N, Lowry S, Mock S, Helmke M, Kubach A, Narro M, Hopkins N, Micklos D, Hilgert U, Gonzales M, Jordan C, Skidmore E, Dooley R, Cazes J, McLay R, Lu Z, Pasternak S, Koesterke L, Piel WH, Grene R, Noutsos C, Gendler K, Feng X, Tang C, Lent M, Kim S-J, Kvilekval K, Manjunath BS, Tannen V, Stamatakis A, Sanderson M, Welch SM, Cranston K, Soltis P, Soltis D, O'Meara B, Ane C, Brutnell T, Kleibenstein DJ, White JW, Leebens-Mack J, Donoghue MJ, Spalding EP, Vision TJ, Myers CR, Lowenthal D, Enquist BJ, Boyle B, Akoglu A, Andrews G, Ram S, Ware D, Stein L, Stanzione D (2011) The iPlant Collaborative: cyberinfrastructure for plant biology. Front Plant Sci 2 (34). doi:10.3389/fpls.2011.00034
13. Goecks J, Nekrutenko A, Taylor J, Team TG (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86. doi:10.1186/gb-2010-11-8-r86
14. Loraine AE, McCormick S, Estrada A, Patel K, Qin P (2013) RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. Plant Physiol 162(2):1092–1109. doi:10.1104/pp.112.211441
15. Biswas S, Agrawal YN, Mucyn TS, Dangl JL, Jones CD (2013) Biological averaging in RNA-Seq. arXiv 1309.0670
16. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15(2):121–132. doi:10.1038/nrg3642
17. Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7(2): e30619
18. 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
19. 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
20. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31 (1):46–53. doi:10.1038/nbt.2450
21. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33 (3):290–295
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28 (5):511–515. doi:10.1038/nbt.1621
23. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14 (2):178–192. doi:10.1093/bib/bbs017
24. Foissac S, Sammeth M (2015) Analysis of alternative splicing events in custom gene datasets by AStalavista. Methods Mol Biol 1269:379–392. doi:10.1007/978-1-4939-2291-8_24
25. Foissac S, Sammeth M (2007) ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. Nucleic Acids Res 35(suppl 2):W297–W299. doi:10.1093/nar/gkm311
26. Brooks AN, Yang L, Duff MO, Hansen KD, Park JW, Dudoit S, Brenner SE, Graveley BR (2011) Conservation of an RNA regulatory map between Drosophila and mammals. Genome Res 21(2):193–202
27. Katz Y, Wang ET, Stilterra J, Schwartz S, Wong B, Thorvaldsdóttir H, Robinson JT, Mesirov JP, Airoldi EM, Burge CB (2014) Sashimi plots: quantitative visualization of alternative isoform expression from RNA-seq data. arXiv:1306.3466. doi:10.1101/002576

Chapter 8

Information Resources for Functional Genomics Studies in *Brachypodium distachyon*

Keiichi Mochida and Kazuo Shinozaki

Abstract

Online tools and databases play an essential role in the promotion of functional genomics studies. Several resources for information regarding *Brachypodium distachyon* (Brachypodium) are available on the Web. In this chapter, we focus on recently published resources for Brachypodium research. The Brachypodium.org website (http://www.brachypodium.org/) is an information portal that provides links to various genomic resources regarding Brachypodium, including genome annotation and re-sequencing datasets of accessions. RIKEN Full-length cDNA Database (RBFLDB, http://brachy.bmep.riken.jp/ver.1/index.pl) is a web-accessible database that provides information of Brachypodium full-length cDNAs (FLcDNAs) collected in RIKEN and updated gene structures of Brachypodium based on the FLcDNA sequences as well as results of comparative analyses with available sequence resources for *Triticeae* crops, wheat, and barley. We introduce the functionalities and availability of these important information resources. Furthermore, we also present brief descriptions of useful online tools that facilitate Brachypodium functional genomics studies.

Key words Brachypodium, Database, Genome annotation, Full-length cDNA, Online tools

1 Introduction

Information resources on the Internet play a prominent role in modern life sciences. They have facilitated access to biological datasets and bio-resources, thus benefiting the heuristic approach of researchers. Recent high-throughput technological advances have enabled the development of sequence-based resource collections and related resource platforms for specific organisms [1]. Therefore, the role of information resources becomes even more important for functional genomics, not only in well-established model organisms but also in newly emerging ones [2].

Brachypodium distachyon (Brachypodium) was proposed as a model plant by Draper et al. in 2001 to elucidate the biological systems in temperate grasses, cool-season cereals, and crops for biofuel production [3]. Since this plant belongs to the *Pooideae* subfamily that includes major crops such as wheat, barley, rye, and

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_8,

oats, Brachypodium serves as a model plant to promote gene discovery studies for improvements of these crops [4]. In the context of improvement of cellulosic biomass productivity in biofuel crops, Brachypodium provides a useful model system to elucidate molecular systems, particularly in grass plant cell wall. After the whole-genome sequence of the inbred line Bd21 was deciphered, which provided a high-quality reference genome sequence [5], Brachypodium was widely accepted as a model grass and several efforts are underway to develop the genomic resources for this plant [6–8].

The web resources that function as portal sites for each organism are essential to integrate discrete datasets as well as to promote activities of the research community. Several sites providing information regarding model organisms, such as Mouse Genome Informatics (MGI) for mouse (http://www.informatics.jax.org/) [9], FlyBase for the common fly (http://flybase.org/) [10], WormBase for *Caenorhabditis elegans* (https://www.wormbase.org/) [11], and The Arabidopsis Information Resource (TAIR) for Arabidopsis (https://www.arabidopsis.org/) [12] provide information regarding the whole-genome sequence of each organism. Recent innovations in DNA sequencing technologies have provided genome-wide sequence information relatively rapidly, thus increasing the significance of information resources with portal functions for easy access to genomic resources and online tools.

Full-length cDNA (FLcDNA) libraries and large-scale sequence data of those clones are also essential genomic resources for promoting functional genomics studies in model organisms. FLcDNA sequence resources are useful for accurate identification of genomic structural contexts such as transcription units, transcription start sites (TSSs), and transcriptional variants [13–15]. Therefore, resources containing FLcDNA information of a genome-sequenced organism often provide transcription unit structures based on FLcDNA and associated information [16–19].

In this chapter, we describe recently published information resources for *B. distachyon* research, especially two useful resources, namely the Brachypodium.org website (http://www.brachypodium.org/) and the RIKEN Brachypodium full-length cDNA database (RBFLDB; http://brachy.bmep.riken.jp/ver.1/index.pl). Furthermore, we briefly introduce other useful web-accessible tools and databases for Brachypodium studies.

2 Methods

2.1 The Brachypodium.org Website

Brachypodium.org (http://brachypodium.org/) is a portal for Brachypodium genomics that is publicly available, and it enables access to the reference genome annotation of Brachypodium Bd21, whole-genome re-sequencing datasets of *B. distraction* accessions,

genome browser, and services for sequence-similarity search (Fig. 1). The site is accessible using general web browsers such as Microsoft Internet Explorer and Safari, based on http protocol (http://brachypodium.org/), as well as computer applications for FTP clients, based on ftp protocol (ftp://brachypodium.org/brachypodium.org/). The homepage of the site provides a search interface for keywords corresponding to entries in the database. Data are maintained by the Mockler Lab at the Donald Danforth Plant Science Center.

2.2 DB Link on the Brachypodium.org Website

Users can access information regarding gene annotation of the Bd21 genome from the "DB" link on the Brachypodium website, which links to an MIPS v1.2 annotation (http://mips.helmholtz-muenchen.de/plant/brachypodium/download/index), containing 26,552 gene loci coding for 31,029 distinct mRNA molecules. Clicking the hyperlink "Browse Contigs" on the DB page allows for navigation to the websites showing images of each chromosome and scaffold, which provide hyperlinks to the list of genes annotated on them. Using text boxes for start and stop on the page, users can extract a sequence between the nucleotide positions in a FASTA format file. The text box for the search is used to identify or filter the gene list, where each gene name is hyperlinked to details of the corresponding gene on a separate page, which provides the exon–intron structure of annotated gene models, domain organization of deduced protein sequences, and functional annotations based on the Gene Ontology.

2.3 FTP Link on the Brachypodium.org Website

The "FTP" link provides access to various large-scale datasets that are downloadable via an Internet browser. The FTP site is also accessible using FTP clients with anonymous user logins. The FTP contains seven directories: Affymetrix, Annotation, Assembly, Bd21 Original RNAseq, ESTs, Natural variation, and Stress. The "Affymetrix" directory includes library files for a Brachypodium Affymetrix GeneChip, such as Chip Description File (.cdf), Binary Probe Map file (.bpmap), and sequences of the designed probes. The "Annotation" directory includes gff files for gene structural annotation and sequence datasets of promoter regions based on the MIPS v1.2 annotation. The "Assembly" directory includes the FASTA format file of the whole-genome sequence of the Bd21 genome corresponding to the MIPS v1.2 annotation. The "Bd21_Original_RNAseq" directory includes six FASTQ files. These are RNA-Seq-based transcriptome datasets using Illumina sequencing, generated to predict gene structure for the genome sequencing project of Bd21. The "EST" directory provides FASTA files of Brachypodium ESTs from Bd21 and some other accessions based on the Sanger sequencing. The "NaturalVariation" directory provides results of re-sequencing analysis of six Brachypodium accessions (Bd1-1, Bd21-3, Bd3-1, Bd30-1, BdTR12c, and

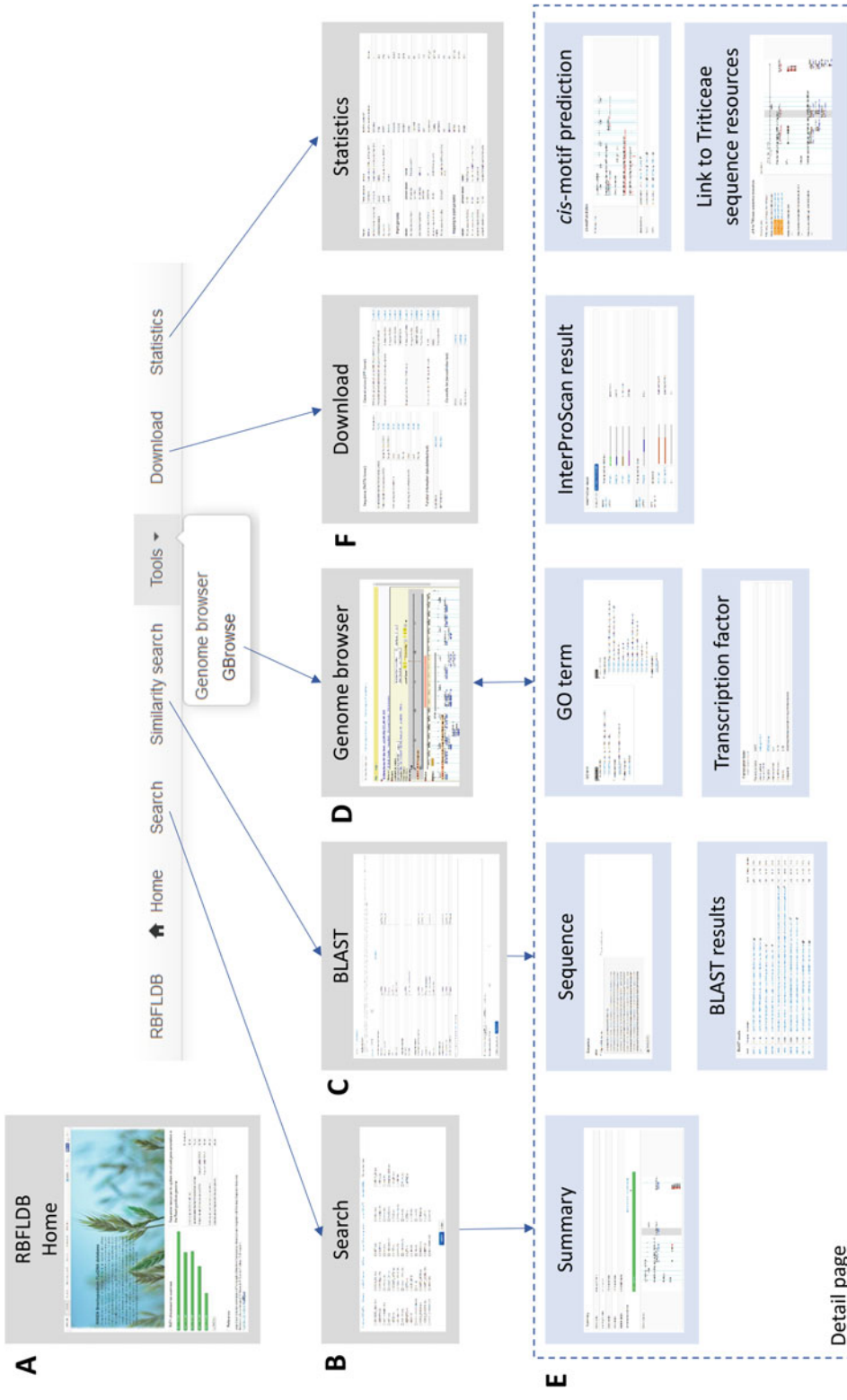

Fig. 1 User interface of the Brachypodium.org website. The top page of the Brachypodium.org website (Home) displays links to each site (**a**). The DB page provides access to gene annotation of each gene model (**b**). The detailed annotation pages provide summarized basic information on each of the gene models annotated with gene structure and predicted gene function (**c**). Users also are able to access to FTP site (**d**). Jbrowse is available to browse various annotation allocated on the genome (**e**). BLAST and BLAT are available to search for sequences based on sequence similarities (**f**)

Koz-3 and Bd21 as the control), including results of the identification of polymorphisms such as SNPs, CNV, and chromosomal structural variations.

2.4 JBrowse on the Brachypodium.org Website

The Brachypodiu.org website uses JBrowse as the genome browser to enable users to explore annotated gene structures and associated information on the *B. distachyon* reference genome. Users can select tracks corresponding to each dataset. Using left-click and right-click on tDNA, the set of lines containing the tDNA and the tDNA ordering website can be accessed, respectively. For optimal results, it is recommended that JBrowse be used with IE9 and Safari.

2.5 BLAST and BLAT on the Brachypodium.org Website

For web-based sequence similarity search, Brachypodium.org provides user interfaces for BLAST and BLAT programs. The BLAST and BLAT interface allows users to query for nucleotide or amino acid sequences against the Bd21 genome sequence and CDS of annotated transcription units and deduced proteome datasets in the Bd21 genome. Users can also select ESTs from Bd21 and some other accessions in the search using the BLASTN program. The user interface of a search result includes hyperlinks to genome or gene models that navigate users to the page containing more details of the corresponding gene.

2.6 RIKEN Brachypodium Full-Length cDNA Database (RBFLDB)

RBFLDB (http://brachy.bmep.riken.jp/ver.1/index.pl) is a web-accessible database that provides information related to Brachypodium full-length cDNAs collected in RIKEN (Fig. 2) [20]. The full-length cDNA library was constructed from a mixed RNA sample from 21 various tissues. Using the Sanger sequencing method, 78,163 high-quality expressed sequence tags (ESTs) of ca. 40,000 clones and 16,079 contigs of entirely sequenced clones were generated. This sequence information was used to update gene structure annotations of genes on the *B. distachyon* Bd21 genome. As a result, ca. 10,000 non-redundant gene models were re-annotated by the full-length cDNAs and ca. 6000 gene models were modified in terms of their transcription units especially in UTRs. The database provides various genomic properties of Brachypodium genes updated by using the RBFL cDNA sequences such as gene functions, gene structures, deduced protein domain, and cis-motifs in putative promoter regions. Additionally, the RBFLDB also provides results of comparative analysis with the genomic framework of barley and full-length cDNAs information of wheat and/or barley provided by TriFLDB (http://trifldb.psc.riken.jp).

2.7 Search Page in RBFLDB

The search page in the RBFLDB website provides seven tabs for different types of search queries: IDs (Gene/EST/HTC), Keyword, InterProScan result, GO Terms, cis-motif (Stress responsive), some cis-motif search interfaces (cis-motif (PLACE),

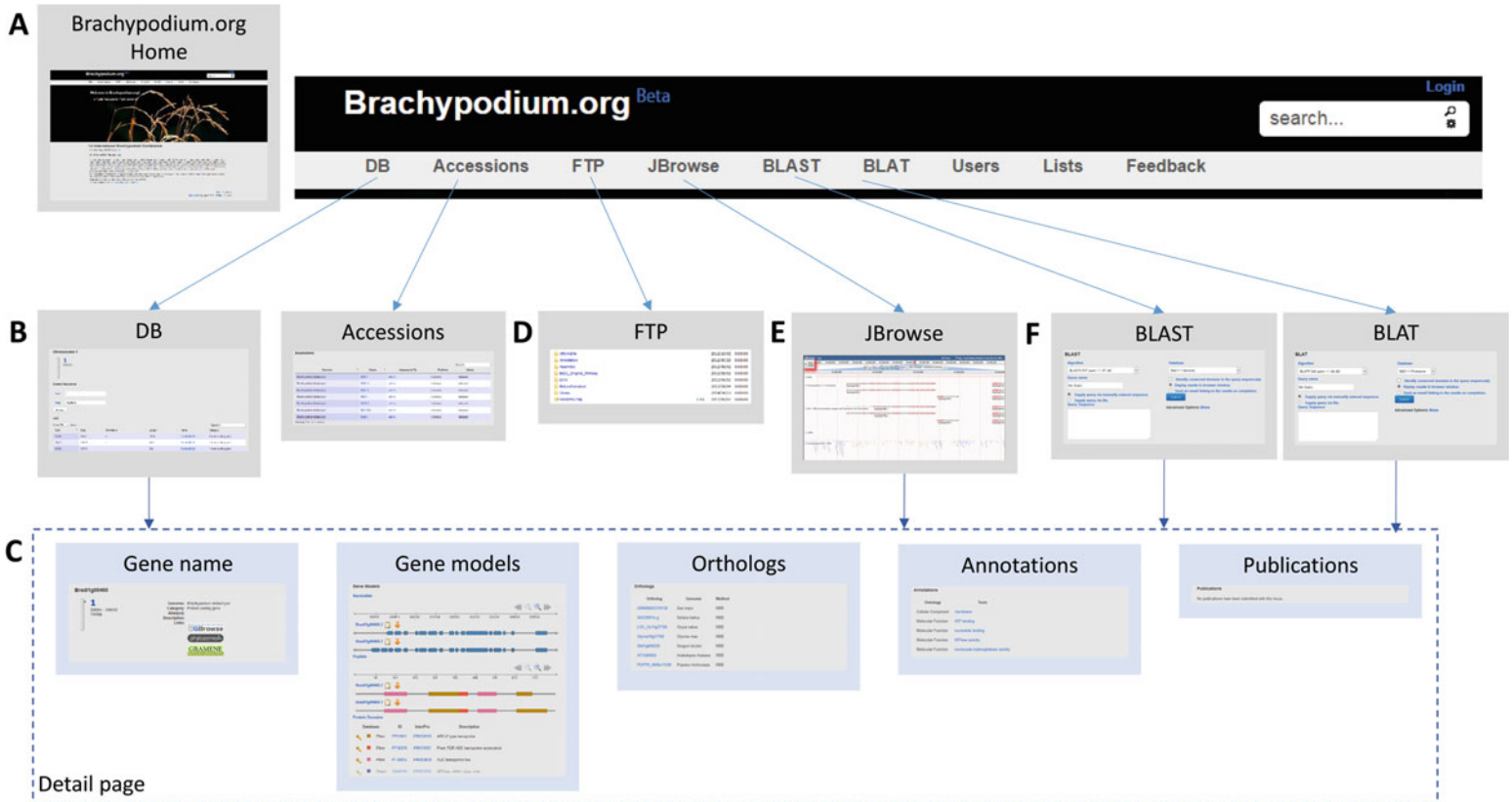

Fig. 2 User interface of the RBFLDB website. The top page of the RBFLDB (Home) displays links to each site (**a**). From the "Search" link, users will be directed to the search page which provides search interface for keywords, sequence identifiers, identifiers of domains found by InterProScan, GO terms, and available cis-motifs (**b**). BLAST is available to search for sequences based on sequence similarities (**c**). Gbrowse is available to browse various annotation allocated on the genome (**d**). The detailed annotation pages provide summarized basic information on each of the gene models annotated with updated gene structure based on the FLcDNA sequences and predicted gene function as well as comparative analysis results with wheat and barley (**e**). Users also are able to download datasets of FLcDNAs and associated information of the RBFLDB (**f**)

cis-motif (AGRIS)), and Transcription factors. The "IDs (Gene / EST/HTC)" search page provides an interface to search full-length cDNA entries by using an identifier of each gene (Bradi*g*****) based on the MIPS v1.2 annotation or the accession number of each RBFL EST or each RBFL HTC. The "Keyword" search page provides an interface to search for entries that include keywords in a string associated with each entry. The definition strings of the putative homologs found in each sequence database have been assembled as a keyword database. The keyword database of homologs is used to search and retrieve relevant gene models corresponding to user's queries. The entries in the RBFLDB were also searched against the InterPro database. Using the "InterProScan result" search interface, users can search for the RBFLDB entries of protein sequences that contain conserved domains found by iprscan search. The "GO Terms" search interface can be used to search for the RBFLDB entries annotated with the gene ontology (GO) terms by InterProScan or by a homology search against the *Arabidopsis* genes with the GO Terms in the TAIR10 database. The "*cis*-motif" search interfaces can be used to search for stress-responsive *cis*-motifs in the promoter regions of all entries in RBFLDB. Users can identify genes whose promoter regions contain each motif as *cis*-element(s) of their interest. The promoter regions of the RBFLDB genes used in the search are defined as upstream sequences −500, −1000, and −3000 bp from the start site of the "gene" or "mRNA" of the updated annotations with the FLcDNAs. Finally, the "Transcription factor" search page can be used to search for gene models that encode putative transcription factors in *Brachypodium*, which were contained in GramineaeTFDB (http://gramineaetfdb.psc.riken.jp/help.pl) [21].

2.8 Similarity Search Interface in RBFLDB

RBFLDB provides "Similarity search" interface for sequence similarity searches using the NCBI BLAST program to search for RBFL entries together with sequences of other plant species. The BLAST program provides cDNA and protein databases of *B. distachyon* as well as of wheat, barley, and others such as rice and Arabidopsis.

2.9 Genome Browser in RBFLDB

The RBFLDB website uses GBrowse as the genome browser to enable users to explore the structural annotations of updated genes with the full-length cDNAs. GBrowse also provides genome browsers of the barley genomic framework with comparative mapping results of Brachypodium genes on to the barley genome.

2.10 Detail Page for Each Gene Model in RBFLDB

Users can view detailed information on full-length cDNA clones, gene annotation (including gene structure, cDNA, and protein sequences of corresponding gene models), protein domain structures predicted by InterProScan, motifs found in promoter regions and GO terms derived from InterProScan and from putative homologs of Arabidopsis. Results of the similarity search against various

sequence databases are summarized on the detail page. It should be useful to predict gene function based on those of the most likely counterparts found in other databases. Information on the cis-motifs located in the promoter region of each gene model is also provided in the detailed page. Furthermore, the detail page also includes a user interface "Link to Triticeae sequence resources" to provide orthologous relationships between Brachypodium and wheat or barley. The interface also provides hyperlinks to Triticeae Full-length cDNA Database (TriFLDB, http://trifldb.psc.riken.jp/v3/index.pl) [22].

2.11 Download Links in RBFLDB

Users can download datasets housed in RBFLDB from the "Download" page as FASTA-formatted sequence datasets, contigs, and ESTs of the RBFLcDNAs and cDNA, CDS, and peptide sequences, based on the updated gene structure based on the RBFLcDNAs. The updated gene structure, mapping results of RBFLcDNAs to the Brachypodium Bd21 genome and the barley genomic framework and allocation of cis-motifs searched in −3000 bp promoter regions as a GFF formatted text file.

2.12 Working Example to Search for Brachypodium cDNAs in RBFLDB

By means of the various types of user interfaces of the RBFLDB website, users can search for sequences and clones of cDNAs corresponding to the *B. distachyon* genes, which will be useful in further experiments to elucidate gene functions and in comparative analyses to understand their evolutionary properties. Users can search for sequences of interest on the search page and the Web BLAST interface, and browse their functional and structural annotations on the detail page for each gene and on the genome browser, respectively. These annotations can provide useful information for further specific applications, such as those described below.

2.13 Isolation of Sequences of Genes and Putative Promoter Regions

Structural gene annotation based on full-length cDNAs should be particularly helpful to identify transcription start sites and promoter regions. Sequences of putative promoter regions can be browsed by using the genome browser, Gbrowse. Users can design PCR primers to amplify a promoter region, and clone the amplified sequences to be used in vector construction for functional analyses of genes such as analysis of promoter activity and regulation of gene expression.

2.14 Search and Order Full-Length cDNA Clones of B. distachyon

The detail page of each *Brachypodium* gene contains accession numbers of ESTs corresponding to the full-length cDNA clones of *B. distachyon*. Users can refer to these accession numbers to search and order for the cDNA clones on the website of RIKEN Bioresource Center (http://epd.brc.riken.jp/en/pdna/pdb_distachyon).

2.15 Search Orthologs Between Brachypodium and Triticeae Crops

The detail page of each *Brachypodium* gene lists hyperlinks that lead to putative counterparts of genes in wheat and barley that are annotated in TriFLDB. Users can also browse comparative mapping results between each of Brachypodium genes and its homologs in wheat and barley.

2.15.1 Other Useful Information Resources for Brachypodium Research

Here, we also present brief descriptions of other useful online resources that could facilitate functional genomics studies in *Brachypodium*. The name and URL of each website and a brief description are summarized in Table 1. The list is an updated version of the "Online resources for *Brachypodium* research" published by Mochida and Shinozaki (2013) [8].

Table 1
Information resources for *Brachypodium* research

	Descriptions	URLs	Ref.
Phytozome 10	• Genome sequence • Genome annotation	http://phytozome.jgi.doe.gov/pz/portal.html	
Brachypodium genome database (MIPS)	• Genome sequence • Genome annotation • Domain search results • Upstream and downstream sequences of each gene model • Comparative map view	http://pgsb.helmholtz-muenchen.de/plant/brachypodium/index.jsp	[23]
NCBI Assembly	• Genome sequence • Genbank accessions and Refseq accessions	http://www.ncbi.nlm.nih.gov/assembly/GCF_000005505.1/	
Ensembl Plants	• Genome sequence • Genome annotation • Wheat RNA-Seq, EST and UniGene alignments	http://plants.ensembl.org/Brachypodium_distachyon/Info/Index	
Gramene	• Gene annotation • Comparative maps • Pathways (BrachyCyc)	http://www.gramene.org	[24]
PlantGDB	• Genome sequence • Genome annotation • Transcripts assembly	http://www.plantgdb.org/	[25]
BRACHYTIL	• TILLING mutant collection	http://urgv.evry.inra.fr/UTILLdb	
Western Regional Research Center	• T-DNA insertional mutant collection	http://brachypodium.pw.usda.gov/	[26]
GramineaeTFDB	• Transcription factors	http://gramineaetfdb.psc.riken.jp/	[21]

(continued)

Table 1
(continued)

	Descriptions	URLs	Ref.
PlantTFDB	• Transcription factors	http://planttfdb.cbi.pku.edu.cn	[27]
BrachyCyc	• Predicted metabolic pathways	http://www.gramene.org/node/126	
KEGG	• Predicted metabolic pathways • Predicted gene function	http://www.genome.jp/kegg-bin/show_organism?org = bdi	
MapMan	• Mapping to MapMan Ontology	http://mapman.gabipd.org/web/guest	
Plant Genome Duplication Database	• Syntenic relationships • Gene annotation • Gene Ontology	http://chibba.agtec.uga.edu/duplication/	[28]
PLAZA	• Syntenic relationships	http://bioinformatics.psb.ugent.be/plaza/	[29]
PlaNet	• Gene expression profiles based on the Brachypodium GeneChip	http://aranet.mpimp-golm.mpg.de/	[30]
E-TALEN	• Web tool to design TALENs	http://www.e-talen.org/E-TALEN/	[31]
CRISPR-P	• Web tool to CRISPR design	http://cbi.hzau.edu.cn/crispr/	[32]
Jaiswal Lab at Oregon State University	• Resources of *B. sylvaticum*	http://jaiswallab.cgrb.oregonstate.edu/genomics/brasy	
PIGD	• Intronless genes in the Poaceae	http://pigd.ahau.edu.cn/	[33]
Wheat Zapper	• Prediction of orthologous relationships with the model species including Brachypodium	http://wge.ndsu.nodak.edu/wheatzapper/	[34]
PlantRNA	• Database for tRNAs in plants	http://plantrna.ibmp.cnrs.fr/	[35]
Brachypedia	• Growth stages	http://brachypedia.bmep.riken.jp/wiki/index.php/Image_library	[36]
PRIMe DROP Met	• Metabolome datasets	http://prime.psc.riken.jp/?action=drop_index	[37]

Acknowledgements

This work was partially supported by the Advanced Low Carbon Technology Research and Development Program (ALCA, J2013403) from the Japan Science and Technology Agency (JST).

References

1. Mochida K, Shinozaki K (2010) Genomics and bioinformatics resources for crop improvement. Plant Cell Physiol 51(4):497–523. doi:10.1093/pcp/pcq027
2. Mochida K, Shinozaki K (2011) Advances in omics and bioinformatics tools for systems analyses of plant functions. Plant Cell Physiol 52 (12):2017–2038. doi:10.1093/pcp/pcr153
3. Draper J, Mur LA, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge AP (2001) Brachypodium distachyon. A new model system for functional genomics in grasses. Plant Physiol 127(4):1539–1555
4. Bevan MW, Garvin DF, Vogel JP (2010) Brachypodium distachyon genomics for sustainable food and fuel production. Curr Opin Biotechnol 21(2):211–217. doi:10.1016/j.copbio.2010.03.006
5. International Brachypodium I (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463 (7282):763–768. doi:10.1038/nature08747
6. Brkljacic J, Grotewold E, Scholl R, Mockler T, Garvin DF, Vain P, Brutnell T, Sibout R, Bevan M, Budak H, Caicedo AL, Gao C, Gu Y, Hazen SP, Holt BF 3rd, Hong SY, Jordan M, Manzaneda AJ, Mitchell-Olds T, Mochida K, Mur LA, Park CM, Sedbrook J, Watt M, Zheng SJ, Vogel JP (2011) Brachypodium as a model for the grasses: today and the future. Plant Physiol 157(1):3–13. doi:10.1104/pp.111.179531
7. Mur LA, Allainguillaume J, Catalan P, Hasterok R, Jenkins G, Lesniewska K, Thomas I, Vogel J (2011) Exploiting the Brachypodium tool box in cereal and grass research. New Phytol 191(2):334–347. doi:10.1111/j.1469-8137.2011.03748.x
8. Mochida K, Shinozaki K (2013) Unlocking Triticeae genomics to sustainably feed the future. Plant Cell Physiol 54(12):1931–1950. doi:10.1093/pcp/pct163
9. Shaw DR (2009) Searching the mouse genome informatics (MGI) resources for information on mouse biology from genotype to phenotype. Curr Protoc Bioinformatics/editorial board, Andreas D Baxevanis [et al] Chapter 1: Unit1 7. doi:10.1002/0471250953.bi0107s25
10. dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase C (2015) FlyBase: introduction of the Drosophila melanogaster release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res 43(Database issue): D690–D697. doi:10.1093/nar/gku1099
11. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, de la Cruz N, Duong A, Fang R, Ganesan U, Grove C, Howe K, Kadam S, Kishore R, Lee R, Li Y, Muller HM, Nakamura C, Nash B, Ozersky P, Paulini M, Raciti D, Rangarajan A, Schindelman G, Shi X, Schwarz EM, Ann Tuli M, Van Auken K, Wang D, Wang X, Williams G, Hodgkin J, Berriman M, Durbin R, Kersey P, Spieth J, Stein L, Sternberg PW (2012) WormBase 2012: more genomes, more data, new website. Nucleic Acids Res 40(Database issue): D735–D741. doi:10.1093/nar/gkr954
12. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The Arabidopsis information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40(Database issue):D1202–D1210. doi:10.1093/nar/gkr1090
13. Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, Muramatsu M, Hayashizaki Y, Kawai J, Carninci P, Itoh M, Ishii Y, Arakawa T, Shibata K, Shinagawa A, Shinozaki K (2002) Functional annotation of a full-length Arabidopsis cDNA collection. Science 296 (5565):141–145. doi:10.1126/science.1071006
14. Iida K, Seki M, Sakurai T, Satou M, Akiyama K, Toyoda T, Konagaya A, Shinozaki K (2004) Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences. Nucleic Acids Res 32(17):5096–5103. doi:10.1093/nar/gkh845
15. Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J (2009)

Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. Plant J 60(2):350–362. doi:10.1111/j.1365-313X.2009.03958.x

16. Akiyama K, Kurotani A, Iida K, Kuromori T, Shinozaki K, Sakurai T (2014) RARGE II: an integrated phenotype database of Arabidopsis mutant traits using a controlled vocabulary. Plant Cell Physiol 55(1):e4. doi:10.1093/pcp/pct165
17. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, Wakimoto H, Yang CC, Iwamoto M, Abe T, Yamada Y, Muto A, Inokuchi H, Ikemura T, Matsumoto T, Sasaki T, Itoh T (2013) Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 54(2):e6. doi:10.1093/pcp/pcs183
18. Imanishi T, Itoh T, Suzuki Y, O'Donovan C, Fukuchi S, Koyanagi KO, Barrero RA, Tamura T, Yamaguchi-Kabata Y, Tanino M, Yura K, Miyazaki S, Ikeo K, Homma K, Kasprzyk A, Nishikawa T, Hirakawa M, Thierry-Mieg J, Thierry-Mieg D, Ashurst J, Jia L, Nakao M, Thomas MA, Mulder N, Karavidopoulou Y, Jin L, Kim S, Yasuda T, Lenhard B, Eveno E, Suzuki Y, Yamasaki C, Takeda J, Gough C, Hilton P, Fujii Y, Sakai H, Tanaka S, Amid C, Bellgard M, Bonaldo Mde F, Bono H, Bromberg SK, Brookes AJ, Bruford E, Carninci P, Chelala C, Couillault C, de Souza SJ, Debily MA, Devignes MD, Dubchak I, Endo T, Estreicher A, Eyras E, Fukami-Kobayashi K, Gopinath GR, Graudens E, Hahn Y, Han M, Han ZG, Hanada K, Hanaoka H, Harada E, Hashimoto K, Hinz U, Hirai M, Hishiki T, Hopkinson I, Imbeaud S, Inoko H, Kanapin A, Kaneko Y, Kasukawa T, Kelso J, Kersey P, Kikuno R, Kimura K, Korn B, Kuryshev V, Makalowska I, Makino T, Mano S, Mariage-Samson R, Mashima J, Matsuda H, Mewes HW, Minoshima S, Nagai K, Nagasaki H, Nagata N, Nigam R, Ogasawara O, Ohara O, Ohtsubo M, Okada N, Okido T, Oota S, Ota M, Ota T, Otsuki T, Piatier-Tonneau D, Poustka A, Ren SX, Saitou N, Sakai K, Sakamoto S, Sakate R, Schupp I, Servant F, Sherry S, Shiba R, Shimizu N, Shimoyama M, Simpson AJ, Soares B, Steward C, Suwa M, Suzuki M, Takahashi A, Tamiya G, Tanaka H, Taylor T, Terwilliger JD, Unneberg P, Veeramachaneni V, Watanabe S, Wilming L, Yasuda N, Yoo HS, Stodolsky M, Makalowski W, Go M, Nakai K, Takagi T, Kanehisa M, Sakaki Y, Quackenbush J, Okazaki Y, Hayashizaki Y, Hide W, Chakraborty R, Nishikawa K, Sugawara H, Tateno Y, Chen Z, Oishi M, Tonellato P, Apweiler R, Okubo K, Wagner L, Wiemann S, Strausberg RL, Isogai T, Auffray C, Nomura N, Gojobori T, Sugano S (2004) Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol 2(6):e162. doi:10.1371/journal.pbio.0020162
19. Bono H, Kasukawa T, Furuno M, Hayashizaki Y, Okazaki Y (2002) FANTOM DB: database of functional annotation of RIKEN mouse cDNA clones. Nucleic Acids Res 30(1):116–118
20. Mochida K, Uehara-Yamaguchi Y, Takahashi F, Yoshida T, Sakurai T, Shinozaki K (2013) Large-scale collection and analysis of full-length cDNAs from Brachypodium distachyon and integration with Pooideae sequence resources. PLoS One 8(10):e75265. doi:10.1371/journal.pone.0075265
21. Mochida K, Yoshida T, Sakurai T, Yamaguchi-Shinozaki K, Shinozaki K, Tran LS (2011) In silico analysis of transcription factor repertoires and prediction of stress-responsive transcription factors from six major gramineae plants. DNA Res 18(5):321–332. doi:10.1093/dnares/dsr019
22. Mochida K, Yoshida T, Sakurai T, Ogihara Y, Shinozaki K (2009) TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. Plant Physiol 150(3):1135–1146. doi:10.1104/pp.109.138214
23. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M (2013) MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res 41(Database issue):D1144–D1151. doi:10.1093/nar/gks1153
24. Youens-Clark K, Buckler E, Casstevens T, Chen C, Declerck G, Derwent P, Dharmawardhana P, Jaiswal P, Kersey P, Karthikeyan AS, Lu J, McCouch SR, Ren L, Spooner W, Stein JC, Thomason J, Wei S, Ware D (2011) Gramene database in 2010: updates and extensions. Nucleic Acids Res 39(Database):D1085–D1094. doi:10.1093/nar/gkq1148
25. Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2008) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res 36(Database issue):D959–D965. doi:10.1093/nar/gkm1041
26. Bragg JN, Wu J, Gordon SP, Guttman ME, Thilmony R, Lazo GR, YQ G, Vogel JP (2012) Generation and characterization of the western regional research Center Brachypodium T-DNA insertional mutant collection. PLoS One 7(9):e41916. doi:10.1371/journal.pone.0041916

27. Jin J, Zhang H, Kong L, Gao G, Luo J (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. Nucleic Acids Res 42(Database issue): D1182–D1187. doi:10.1093/nar/gkt1016
28. Lee TH, Tang H, Wang X, Paterson AH (2013) PGDD: a database of gene and genome duplication in plants. Nucleic Acids Res 41 (Database issue):D1152–D1158. doi:10.1093/nar/gks1104
29. Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inze D, Mueller-Roeber B, Vandepoele K (2015) PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Res 43 (Database issue):D974–D981. doi:10.1093/nar/gku986
30. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S (2011) PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell 23(3):895–910. doi:10.1105/tpc.111.083667
31. Heigwer F, Kerr G, Walther N, Glaeser K, Pelz O, Breinig M, Boutros M (2013) E-TALEN: a web tool to design TALENs for genome engineering. Nucleic Acids Res 41(20):e190. doi:10.1093/nar/gkt789
32. Lei Y, Lu L, Liu HY, Li S, Xing F, Chen LL (2014) CRISPR-P: a web tool for synthetic single-guide RNA design of CRISPR-system in plants. Mol Plant 7(9):1494–1496. doi:10.1093/mp/ssu044
33. Yan H, Jiang C, Li X, Sheng L, Dong Q, Peng X, Li Q, Zhao Y, Jiang H, Cheng B (2014) PIGD: a database for intronless genes in the Poaceae. BMC Genomics 15:832. doi:10.1186/1471-2164-15-832
34. Alnemer LM, Seetan RI, Bassi FM, Chitraranjan C, Helsene A, Loree P, Goshn SB, YQ G, Luo MC, Iqbal MJ, Lazo GR, Denton AM, Kianian SF (2013) Wheat zapper: a flexible online tool for colinearity studies in grass genomes. Funct Integr Genomics 13(1):11–17. doi:10.1007/s10142-013-0317-4
35. Cognat V, Pawlak G, Duchene AM, Daujat M, Gigant A, Salinas T, Michaud M, Gutmann B, Giege P, Gobert A, Marechal-Drouard L (2013) PlantRNA, a database for tRNAs of photosynthetic eukaryotes. Nucleic Acids Res 41(Database issue):D273–D279. doi:10.1093/nar/gks935
36. Onda Y, Hashimoto K, Yoshida T, Sakurai T, Sawada Y, Hirai M, Toyooka K, Mochida K, Shinozaki K (2015) Determination of growth stages and metabolic profiles in Brachypodium distachyon for comparison of developmental context with Triticeae crops. Proc R Soc Lond B Biol Sci. doi:10.1098/rspb.2015.0964
37. Sakurai T, Yamada Y, Sawada Y, Matsuda F, Akiyama K, Shinozaki K, Hirai MY, Saito K (2013) PRIMe update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation. Plant Cell Physiol 54(2):e5. doi:10.1093/pcp/pcs184

Chapter 9

Methods for Functional Transgenics: Development of Highly Efficient Transformation Protocol in *Brachypodium* and Its Suitability for Advancing *Brachypodium* Transgenics

Ron Vunsh

Abstract

Plant transformation is an invaluable technique in plant genomics by which an extra foreign DNA sequence is introduced into a plant genome. Changing the plant genome is leading to owning new genetic characteristics. Model plant is a keystone in a study of the comprehensive plant phylum. Here, I describe an efficient method to transform the plant species *Brachypodium distachyon* which, due to its characters, is developing to be an important plant model.

Key words *Brachypodium distachyon*, Transformation, *Agrobacterium tumefaciens*, Embryonic callus

1 Introduction

Plant transformation is a technical method by which a small foreign DNA sequence is introduced and integrates into an existing defined genome of the target plant species. This genetical change in a plant genome opened the plant world to new scientific and practical possibilities which can be compared to those which exist in the one cell bacterial world. Many plant species were transformed during the last 50 years [1–3]. Introducing a small DNA sequence into the genome (integration into one of the plant chromosomes) of a plant enables inclusion of new traits in the acceptor plant. Introduction of different genes from different foreign organisms into plant cells, tissues, or organs granting the transformed plant unlimited new characteristics like pest resistance, physical changes, yield increase, etc. [4].

The ability to produce transgenic plants and plant tissues opened a new era in plant practical use as well as in plant research world.

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_9,

Plant transformation has been achieved in many different ways: Bombardment or microinjection of small DNA sequence into a plant cell, using the natural transforming system of *Agrobacterium* and by many other methods.

Transformation efficiency is determined by the relation between the initial challenged quantity of explants and the transformed quantity of plants or tissues acquired. Efficient transformation system is needed in order to use it as a reliable and powerful system in which statistically enough results are gathered scientifically and practically.

Biolistic transformation (developed by Sanford [5]) is applicable [6], yet it is less favorable due to complex transgene insertion patterns containing many copies of the transgene and substantial rearrangements of both the inserted DNA and chromosomal DNA [7]. Microinjection is non-efficient: an individual cell is transformed manually [8]. Agrobacterium-mediated transformation results in low copy number insertions and can be highly efficient [9]. It depends on the physical conditions and the constituents involved in the transformation [10].

Brachypodium distachyon physical, genetic, and genomic attributes make it suitable for use as a model plant system for monocot plants. It is a small plant with simple growth requirements, with short annual life-cycle; it has diploid ecotypes with good self-fertility and its genome is of small size. Homogeneous inbred genotypes and transformation system were developed by J. Vogel [11].

All the cereals, which are the basic and most important food for men and animals are monocots. The phylogenetic relationship between the genus *Brachypodium* and the other grasses revealed that *Brachypodium distachyon* is evolutionary close to the cereals, thus can serve as an ideal model plant for this plant class, performing genetic change induced by transformation [10].

2 Materials

2.1 Plant Material (Table 1)

Compact embryogenic calli derived from immature embryos of *Brachypodium distachyon* accession Bd21-3. Appropriate immature embryos are those of size ~0.3 mm, are bell-shaped, refractive, almost clear as gel (Fig. 1).

Table 1
Chemicals

Material name	Manufacturer name	Catalog number	Storage conditions
Hygromycin B	Duchefa	H0192	2–8 °C
Ticarcillin disodium	Duchefa	T0180	2–8 °C
Carbenicillin disodium	Duchefa	C0109	2–8 °C
Biotin	Duchefa	B0603	2–8 °C
Kanamycin monosulfate	Duchefa	K0126	RT
Casein hydrolysate	Duchefa	C1301	RT
D-mannitol	Duchefa	M0803	RT
Folic acid	Duchefa	F0608	RT
L-Glutamic acid	Duchefa	G0707	RT
Glycine	Duchefa	G0709	RT
Glycerol	Sigma	G7757	RT
K_2HPO_4	Sigma	P9699	RT
LS- salts and vitamins (Linsmaier and Skoog)	Duchefa	C0109	2–8 °C
Maltose monohydrate	Duchefa	M0811	RT
$MgSO_4{\cdot}7H_2O$	Sigma	M188-3006	RT
MS medium + B5 vitamins (Murashige and Skoog + Gamborg's B5 vitamins)	Duchefa	M0231	2–8 °C
NaCl	Duchefa	S0520	RT
Phytagel	Sigma	P8169	RT
Plant agar	Duchefa	P1001	RT
Sucrose	Duchefa	S0809	RT
10% Synperonic PE/F68	Sigma	S81112	RT
Tryptone	Duchefa	T1332	RT
Yeast extract	Conda Pronadisa	1702	RT
LB (Luria Broth)	Duchefa	L1704	RT

RT room temperature

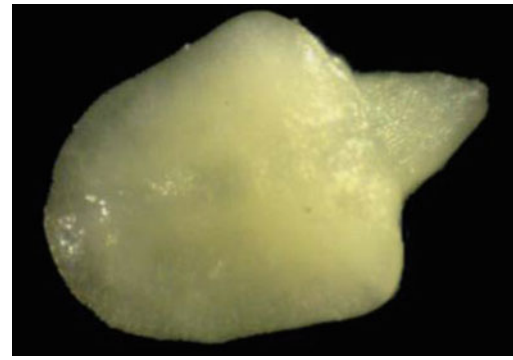

Fig. 1 Embryo ready for transformation (magnification: ×100)

2.2 Bacterial Material

Agrobacterium strain AGL1 containing plasmid pEBh.

1. Prepare *Agrobacterium* glycerol stock: Grow the bacteria in LB medium containing the appropriate antibiotics 10 mL in 50 mL tube in a gyratory shaker. 240 rpm 28 °C overnight in the dark. Pipet 0.85 mL into Eppendorf 1.5 mL tube, add 0.15 mL sterile glycerol, mix and freeze directly in liquid air. Store at −80 °C freezer.
2. Use an *Agrobacterium* glycerol stock tube to grow in LB medium with the appropriate antibiotics for transformation.

2.3 Solution Preparation

2.3.1 Antibiotics: (See **Note 1**)

1. Carbenicillin (200 mg/mL) stock solution, dissolved in Double Distilled Water (DDW).
2. Kanamycin (100 mg/mL) stock solution, dissolved in DDW.
3. HygromycinB (40 mg/mL) stock solution, dissolved in DDW.

2.3.2 General Solutions

1. 6 mg/mL $CuSO_4 \cdot 5H_2O$ stock solution, dissolved in DDW.
2. LB, Bacterial growth medium, dissolved in DDW and autoclaved. Keep in RT.
3. 200 mM Acetosyringone stock solution (*see* **Note 1**)

2.3.3 Hormones (see **Note 2**)

1. 2.5 mg/mL 2,4 Dichlorophenoxy acetic acid (2,4-D) stock solution, dissolved in DDW.
2. 1 mg/mL 6-Benzylaminopurine (BAP) stock solution, dissolved in 1 N NaOH.
3. 1 mg/mL Indol-3-acetic acid (IAA) stock solution, dissolved in 1 N NaOH.
4. 2 mg/mL α-Naphthalene acetic acid (NAA) stock solution, dissolved in ethanol.

Table 2
CIM preculture 0.25 (Callus Induction Medium without selection)

Ingredient	Quantity for 1 L	Final concentration
LS salts and vitamins	4.41 g	0.441%
Sucrose	30 g	3%
Titrate to pH 5.8		
Phytagel	2.5 g	0.25%
After autoclave add:		
CuSO4 (6 mg/mL) stock solution	100 µL	6 mg/l
2,4-D (2.5 mg/mL) stock solution	1000 µL	2.5 mg/l

2.4 Media Preparation (Table 2)

2.4.1 Preparation of "CIM" (Callus Induction Medium) for 1 L

1. Add about 400 mL De-Ionized Water (DIW) into a 1 L Glass Bottle.
2. Add the ingredients while mixing with a magnetic stirrer.
3. Complete the volume with DIW to 1 L using a graduated cylinder.
4. Titrate with 10 M/1 M/0.1 M KOH to reach pH = 5.8.
5. Autoclave the bottles (20 min sterilization 121 °C, 1.2 atm.).
6. Store at room temperature (RT).

2.4.2 Preparation of Solidified CIM 0.25 (Addition of the Phytagel to 500 mL CIM)

1. Transfer 1.25 g Phytagel to a 1 L Glass bottle, add 500 mL of "CIM".
2. Autoclave the bottles (20 min sterilization 121 °C, 1.2 atm.).
3. Store at room temperature (RT).

2.4.3 Preparation of the "CIM Preculture Solid Plates Medium" for 500 mL

1. Dissolve the solidified CIM medium (*see* **Note 3**).
2. Cool the medium to ~60 °C.
3. Take out from the −20 °C freezer 2,4-D (2.5 mg/mL stock solution) tube, and from 4 °C refrigerator $CuSO_4$ (6 mg/mL stock solution) tube. Work under sterile conditions in the laminar flow hood.
4. Pour the medium to 90 mm sterile Petri plates ~30 mL to each.
5. Label "CIM precult 0.25" and leave the plates in the hood until solidified.
6. Wrap the cold medium plates with 1–2 layers of Parafilm (*see* **Note 4**).

Table 3
CIM SL1 (Callus Induction Medium with selection no. 1)

Ingredient	Quantity for 1 L	Final concentration
LS salts and vitamins	4.41 g	0.441%
Sucrose	30 g	3%
Titrate to pH 5.8		
Phytagel	2.5 g	0.25%
After autoclave add:		
$CuSO_4$ (6 mg/mL) stock solution	100 μL	6 mg/l
2,4-D (2.5 mg/mL) stock solution	1000 μL	2.5 mg/l
HygromycinB (40 mg/mL) stock solution	1000 μL	40 mg/l
Ticarcillin (200 mg/mL) stock solution	2000 μL	400 mg/L

2.4.4 CIM Preculture 0.4 (Callus Induction Medium Without Selection, Double Amount of Phytagel)

1. Prepare as CIM preculture 0.25 (2.4.2), but add 2 g. Phytagel into 500 mL CIM medium.
2. Label "CIM precult 0.4" and leave the plates in the hood until solidified (Table 3).

2.4.5 Addition of Solutions and Antibiotics

1. Thaw the Eppendorf tubes of HygromycinB and Ticarcillin (200 mg/mL stock).
2. Work in sterile laminar flow hood. Add 500 μL of HygromycinB (40 mg/mL) and 1000 μL of Ticarcillin (200 mg/mL) to 500 mL medium and shake the bottle for a few seconds.
3. Pour the medium to 90 mm sterile Petri plates (around 17 plates from 500 mL medium), ~30 mL to each.
4. Leave the plates in the hood until solidified.
5. Label the plates with: "CIM SL1" and the preparation date (Table 4).

2.4.6 Prepare SIM SL2 as CIM SL1, with Adjusting the Antibiotics and the Hormones

1. Label the plates with: "SIM SL2" and the preparation date (Table 5).

2.4.7 Prepare SIM3 as CIM SL1, with Adjusting the Antibiotics and the Hormones

1. Label the plates with: "SIM 3" and the preparation date (Table 6).

Table 4
SIM SL2 (Shoot Induction Medium with selection)

Ingredient	Quantity for 1 L	Final concentration
LS salts and vitamins	4.41 g	0.441%
Sucrose	30 g	3%
Titrate to pH 5.8		
Phytagel	2.5 g	0.25%
After autoclave add:		
BAP (1 mg/mL) stock solution	1000 µL	1 mg/l
IAA (1 mg/mL) stock solution	250 µL	0.25 mg/l
HygromycinB (40 mg/mL) stock solution	1000 µL	40 mg/l
Ticarcillin (200 mg/mL) stock solution	2000 µL	400 mg/L

Table 5
SIM3 (shoot induction medium without selection)

Ingredient	Quantity for 1 L	Final concentration
LS salts and vitamins	4.41 g	0.441%
Maltose monohydrate	30 g	3%
Titrate to pH 5.8		
Phytagel	2.5 g	0.25%
After autoclave add:		
BAP (1 mg/mL) stock solution	1000 µL	1 mg/l
IAA (1 mg/mL) stock solution	250 µL	0.25 mg/l
Ticarcillin (200 mg/mL) stock solution	2000 µL	400 mg/L

2.4.8 Prepare RIM as CIM SL1, with Adjusting the Antibiotics and the Hormones

1. Pour the medium into sterile Magenta boxes about 80 mL/box.
2. Label the Magenta box with: "RIM" and the preparation date (Table 7).

Table 6
RIM (Root Induction Medium)

Ingredient	Quantity for 1 L	Final concentration
MS Basal salts	4.31 g	0.431%
Sucrose	30 g	3%
Titrate to pH 5.8		
Phytagel	3.0 g	0.3%
After autoclave add:		
NAA (2 mg/mL) stock solution	1000 µL	2 mg/l
IAA (1 mg/mL) stock solution	1000 µL	1 mg/l
Ticarcillin (200 mg/mL) stock solution	2000 µL	400 mg/L

Table 7
MGL medium (*see* Note 5)

Ingredient	Quantity for 1 L	Final concentration
Tryptone	5.0 g	0.5%
Yeast extract	2.5 g	0.25%
NaCl	5.0 g	0.5%
D-Mannitol	5.0 g	0.5%
$MgSO_4 \cdot 7H_2O$	0.204 g	0.0204%
K_2HPO_4	0.250 g	0.025%
L-Glutamic acid	1.2 g	0.12%
Titrate to pH 7.2		
Plant agar	7.5 g	0.75%
After autoclave add:		
Carbenicillin (200 mg/mL) stock solution	250 µL	50 mg/l
Acetosyringone (200 mM) stock solution	500 µL	100 µM

2.4.9 Preparation of the "Liquid MGL" for 1 L

1. Add about 400 mL DIW into a 1 L Glass Bottle.
2. Add the ingredients while mixing with a magnetic stirrer.
3. Complete the volume with DIW to 1 L.
4. Titrate with 10 M/1 M/0.1 M NaOH to Reach pH = 7.2.

Table 8
Liquid MS for Brachypodium (MS–B)

Ingredient	Quantity for 1 L	Final concentration
MS + B5 vitamins	4.414 g	0.414%
Biotin + glycine + folic acid stock	20 mL	N.A.
Sucrose	30.0 g	3.0%
Casein Hydrolysate	0.8 g	0.08%
Titrate to pH 5.8		

2.4.10 Addition of Plant Agar (for 500 mL)

1. Transfer 3.75 g Plant Agar to a new 1 L Glass bottle, add 500 mL of the "liquid MGL".
2. Autoclave the bottles for 1 h (the regular program: 20 min sterilization at 121 °C, 1.2 atm.).
3. Store at RT.

2.4.11 Preparation of the "MGL Solid Plates Medium" for 500 mL

1. If the medium is solid, melt the Agar in the microwave oven.
2. Cool the sterile medium to ~60 °C using iced water before adding the antibiotics.
3. Thaw the Eppendorf tubes of 200 mg/mL Carbenicillin, stock and 200 mM Acetosyringone at RT.
4. Vortex the solutions in Eppendorf tubes and add 250 μL of Acetosyringone and 175 μL of Carbenicillin to the 500 mL MGL medium.
5. Pour the medium to 90 mm plates (around 17 plates from 500 mL medium) ~30 mL to each.
6. Leave the plates in the laminar hood until solidified.
7. Label the plates with: "MGL" and the preparation date (Table 8).

2.4.12 Preparation of the "Liquid MS-B" for 1 L

1. Add about 400 mL DIW into a 1 L Glass Bottle.
2. Dissolve 5 mg of biotin + 20 mg of glycine + 10 mg of folic acid in 500 mL of DIW and autoclave it (*see* **Note 6**).
3. Add the ingredients while mixing with a magnetic stirrer.
4. Complete the volume with DIW to 1 L.
5. Autoclave the bottles for using the regular program: 20 min sterilization at 121 °C, 1.2 atm.
6. Store at RT.

2.5 Equipment (Tables 9 and 10)

Table 9
Tools and equipment

No.	Equipment	Supplier
1	pH meter	Hanna Instruments
2	Spectrophotometer	Amersham Sciences
3	Gyratory shaker incubator	New Brunswick
4	Pipettors (1–1000 μL)	Labsystems
5	Electronic pipettor (1–25 mL)	Drummond
6	Glass bottle (1 L)	Fisher Brand
7	Balance	Mattler Toledo
8	1000 mL, 100 mL graduated cylinder	Isolab
9	Laminar flow work station (hood)	A.D. Clean rooms
10	Scalpel handle	Swann-Morton
11	Forceps	Medicon
12	Sterilizator (autoclave)	Tuttnauer
13	Dissecting microscope (binocular)	Zeiss
14	−80 °C Freezer	Revco
15	−20 °C Freezer	Amcor
16	Refrigerator	Amcor
17	Centrifuge	Heraeus
18	DDW distiller	Barnstead

3 Methods

3.1 Spike Collection

The spikes are collected from the *Brachypodium* plants approximately 7–10 weeks from sowing. The spikes should be filled with embryos, still soft and flexible when held by fingers. The bottom of the spikelet should feel hard. Ready spikes are removed from the plants by hands keeping the plants intact for upcoming collections. *See* Fig. 2—an immature embryo in a spikelet ready for dissection.

3.1.1 Spikes Sterilization

Fill up to ¾ of 50 mL sterile tube with spikes, add 40 mL 3% bleach and add 40 μL Tween 20, then shake the tube on Gyratory Shaker, for 20 min 80 rpm. Wash the spikes with sterile water three times.

3.1.2 Callus Induction

Place a layer of sterile Whatmann filter paper on a lid of sterile Petri dish.

Table 10
Disposable equipment

No.	Equipment
1	Sterile filter paper (Whatmann No. 1)
2	Sterile pipette tips with filter
3	Sterile plastic tube 50 mL
4	Sterile Petri plates 90 mm
5	Sterile Petri plates 90 mm
6	Sterile Petri plates 140 mm
7	Sterile tissue culture boxes (Magenta)
8	Sterile plastic spreader
9	Sterile cell scraper
10	Silica beads
11	Aluminum foil
12	Parafilm

Fig. 2 Embryo in a spikelet (magnification: ×5)

Put 20–30 spikes on the filter paper to soak excess water. Dissect embryos out of spikes using sterile fine forceps and a dissecting microscope (binocular). The embryo must be plated immediately after isolation on "CIM precult 0.4" medium. After plating all the embryos they must be placed with the scutellum upwards. Place 50–100 embryos on a 90 mm plate. Incubate at 28 °C in the dark. "CEC" (Compact Embryonic Callus) will appear on the embryo 7 days after dissection. A "tale" of shoot and root should be cut-off and the "CEC" is transferred onto a fresh "CIM precult 0.4" medium for 14 days. Transfer the "CEC" to "CIM precult 0.25" 30/plate.

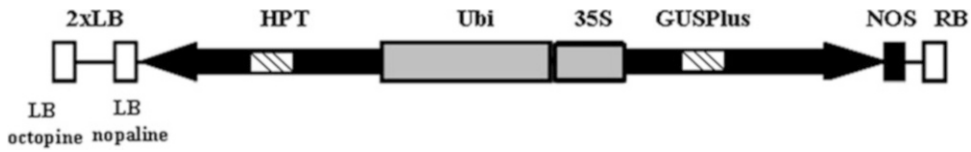

Fig. 3 Diagram of pEBh. Coding sequences are shown as *black* boxes, promoters are *gray* and introns are *hatched*. *LB, RB* left and right T-DNA borders, *Ubi* ubiquitin promoter, *GUS* glucuronidase, *HPT* hygromycin phosphotransferase, *NOS* nopaline synthase terminator

Fig. 4 Scraping the Agrobacteria using a sterile disposable scraper

3.1.3 Transformation

A super-virulent strain of *Agrobacterium*, AGLI is used with the plasmid pEBh (*see* diagram of the plasmid, Fig. 3). Grow the appropriate *Agrobacterium* from a glycerol stock (450 μL stored in −80 °C) and spread it, using Drigalski stick on an MGL plate containing the appropriate antibiotics and 100 μM Acetosyringone. Place the plate upside down at 28 °C. After 24 h incubation (on the day of the transformation), using a sterile scraper, scrape the bacteria from the plate and place it into a 50 mL tube containing 10 mL liquid MS-B (*see* Fig. 4). Pipette the bacteria up and down inside the fluid until the suspension is homogenized. Transfer the homogenized *Agrobacteria* into a 250 mL Erlenmeyer and add 90 mL of MS-B medium. Adjust the Acetosyringone concentration to 200 μM. Shake the Erlenmeyer bottle in RT for about 25 min on Gyratory shaker set at 240 rpm.

Measure the OD_{600} in a spectrophotometer and adjust it to 0.6–1.0 by adding MS-B medium. Add 1 mL of 10% Synperonic to 100 mL of the suspension, shake gently. To a plate with 30 "CEC"

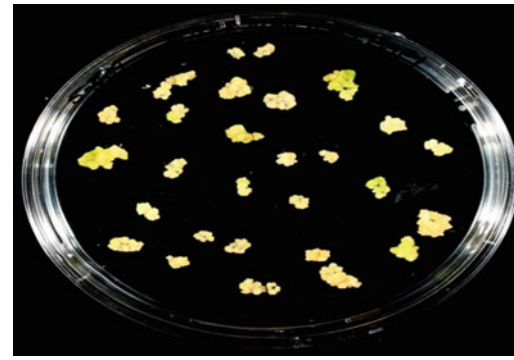

Fig. 5 A plate with Compact Embryonic Calli (CEC) ready for infection

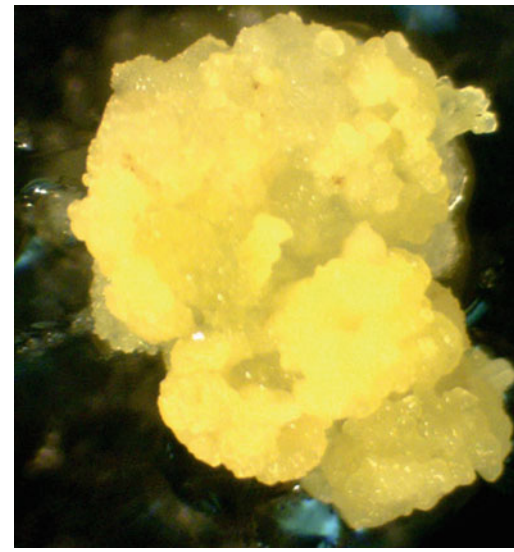

Fig. 6 Compact embryonic callus ready for Agrobacterium infection (magnification: ×10)

add 8.5–10 mL of *Agrobacterium* suspension. Incubate in the hood for 5 min. Pipette the *Agrobacterium* suspension out and transfer the "CEC" on a filter paper. After the "CEC" are dried, transfer them for co-cultivation onto a filter paper placed on top of MS-B Agar medium. Incubate at 24 °C in the dark 3 days (Figs. 5 and 6).

3.1.4 Selection

Transfer the co-cultivated "CEC" (25 per plate) onto plates containing 30 mL of "CIM SL1" medium. Incubate the plates for 2 weeks at 28 °C in the dark. Transfer the calli under binocular onto "SIM SL2" medium choosing the yellow callus leaving the necrotic brown callus. Incubate the "SIM SL-2" plates in the light (12/12 h light/dark cycle) for 10–14 days. After this period transfer onto fresh "SIM SL-2" medium for the same period. At this

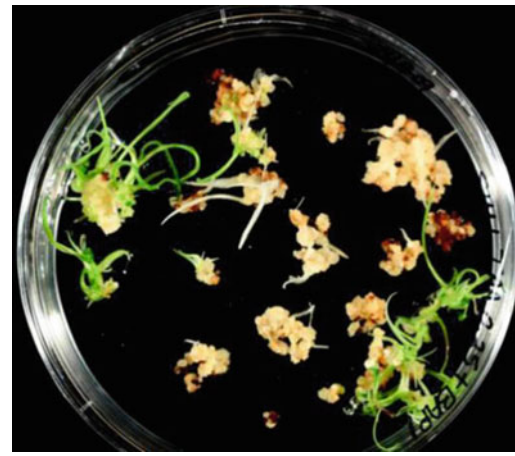

Fig. 7 A plate with transgenic shoot initiations on SL2 medium

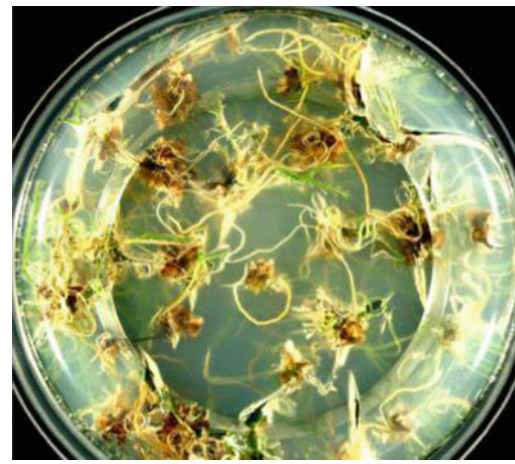

Fig. 8 Roots of regenerated transgenic plants

stage, some of the transgenic calli will turn green and shoot initiations might appear. In another period of 14 days on the same medium, more green calli might appear and the green calli will be developed into small shoots and shoots initiations will begin to grow. When shoots are established they are ready to go into rooting medium. In Fig. 7, young shoot initiations are developed from transgenic calli.

3.1.5 Rooting

Prepare Magenta boxes with RIM medium and insert a transgenic shoot into the "RIM medium." Incubate in a16/8 h. light/dark cycle in 28 °C for 21 days. To those shoots that do not root renew the cut at the bottom of the first leaf and transfer into fresh "RIM medium." Repeat this **steps 1–2** times till you get rooting plantlets. In Fig. 8, typical roots of transgenic plantlets are shown. Plantlets ready for hardening are shown in Fig. 9.

Fig. 9 Rooted transgenic plantlets ready for transfer to hardening in the greenhouse

Fig. 10 Hardened transgenic plants. *Left*: 5 days after transfer to the greenhouse, under a plastic bag. *Right*: 3 weeks in the greenhouse

3.1.6 *Hardening*

Hardening is achieved by transferring the rooted transgenic plants into 7 × 7 × 10 cm pots filled with sterilized mix (peat:perlite: vermiculite, 1:1:1). Planted plants are covered with a plastic bag (with a small opening) and transferred into a greenhouse. During the first week, the plants exposed to the greenhouse environment by withdrawal the plastic bag gradually (Fig. 10).

4 Notes

1. Acetosyringone solution should be fresh. Prepare the solution on the day of using it, by dissolving it in DMSO.
2. Dissolve the antibiotic and the hormones, then sterilize it by filtering through a 0.2 μ filter. Divide it into small portions of 1–2 mL in sterile Eppendorf tubes. Store the sterile tubes in a −20 °C freezer. Thaw before use in 40 °C water bath and vortex it to ensure uniform concentration.
3. Thawing the solidified media can be achieved in microwave oven or in boiling water. After thawing let the media cool to 50–60 °C before adding hormones or antibiotics.
4. All plates with media, with or without plant explants, should be wrapped with 1–2 layers of Parafilm to keep high humidity in the plates and prevent drying of the media.
5. Add to the MGL plate the appropriate antibiotics that will guarantee that only the desirable *Agrobacterium* will grow on the plate.
6. Keep the biotin + glycine + folic acid stock solution in 2–8 °C refrigerator.

Acknowledgments

I gratefully thank Prof. J. Vogel who helped me in my first steps in *Brachypodium distachyon* transformation, supplying the seeds of the Bd21-3 accession and the AGL1 agrobacterium strain.

References

1. Stewart NC, Touraev A, Citovski V (2011) Plant transformation technologies. Wiley-Blackwell, London
2. Wang K (ed) (2006) Agrobacterium protocols Vol. 1 and 2. Humana, New Jersey
3. Jackson JF, Linskens HF (eds) (2013) Genetic transformation of plants. Springer-Verlag, Berlin, Heidelberg
4. Kole C, Michler C, Abbott AG, Hall TC (eds) (2010) Transgenic crop plants: volume 2: utilization and biosafety. Springer-Verlag, Berlin, Heidelberg
5. Sanford CJ (1990) Biolistic plant transformation. Physiol Plant 79:206–209
6. Kikkert JR, Vidal JR, Reisch BI (2005) Chapter 4: Stable transformation of plant cells by particle bombardment/biolistics. In: Pena L (ed) Transgenic plants, methods and protocols. methods in molecular biology, vol 286. Humana, New Jersey
7. Svitashev SK, Somers DA (2002) Characterization of transgene loci in plants using FISH: a picture is worth a thousand words. Plant Cell Tissue Organ Cult 69:205–214
8. Neuhaus G, Spangenberg G (1990) Plant transformation by microinjection techniques. Physiol Plant 79:213–217
9. Travella S, Ross SM, Harden J, Everett C, Snape JW, Harwood WA (2005) A comparison of transgenic barley lines produced by particle bombardment and agrobacterium-mediated techniques. Plant Cell Rep 23:780–789

10. Vogel J, Bragg J (2009) Brachypodium distachyon, a new model for the Triticeae. In: Feuillet C, Muehlbauer GJ (eds) Genetics and genomics of the triticeae, plant genetics and genomics: crops and models 7. Springer, Berlin, pp 427–449

11. Vogel J, Hill T (2008) High-efficiency agrobacterium-mediated transformation of Brachypodium distachyon inbred line Bd21-3. Plant Cell Rep 27:471–478

Chapter 10

Molecular Markers in Whole Genome Evolution of *Brachypodium*

Xin-Chun Mo, De-Quan Zhang, Can Kou, and Ling-Juan Yin

Abstract

Molecular markers play more and more important role in population genetic and phylogenetic studies; choice of marker systems for a particular study has become a serious problem. These marker systems have different advantages and disadvantages, so it is imperative to keep in mind all the pros and cons of the technique while selecting one for the problem to be addressed.

Here, we concisely introduced three molecular marker techniques, namely SSR, ISSR, and RFLP. We elaborated their properties such as reliability, simplicity, cost-effectiveness, and speed, in addition to data analysis of genetic diversity. We have outlined here the whole methodology of these techniques.

Key words Molecular markers, Brachypodium, Whole genome evolution, SSR, ISSR, RFLP

1 Introduction

Molecular markers play a very vital role in taxonomy, physiology, embryology, plant breeding, ecology, genetic engineering, etc., which involved in plant identification and plant improvement [1]. Although whole genome sequence (WGS) is now available for identifying specific genes located on a particular chromosome in some plant species, molecular markers are still an alternative option to study the genetic variation in population level. A genetic marker can be identified by the following ways: (a) a chromosomal landmark or allele that allows for the tracing of a specific region of DNA; (b) a specific piece of DNA with a known position on the genome (Wikipedia-the free encyclopedia; http://en.wikipedia.org/wiki/Genetic_marker); or (c) a gene whose phenotypic expression is usually easily discerned, used to identify an individual or a cell that carries it, or as a probe to mark a nucleus, chromosomes, or locus [2, 3]. They are usually divided into three classes: (1) morphological markers, which depend on the morphological and structural differences; (2) biochemical markers, which are based on gene products (proteins or enzymes); and (3) molecular

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_10,

markers, which depend on a DNA sequence. Early markers like phenotyping and isozymes [4, 5] were few in number and difficult to assay within whole-genome-level analysis (WGA). With quick development of DNA sequencing technology, simple methods based on the DNA sequence were easy to assess genetic variation in the WGA, which could increase sophistication involved in plant evolution and identifying breeding.

This chapter provides a brief description of the methods of molecular marker technology and its application to genus *Brachypodium* Beauv. (Gramineae). This perspective provided a way to assess the genetic variation in WGA of Brachypodium. Molecular marker technology has evolved a series of methods with non-DNA-based methods and DNA-based methods as the technologies for DNA analysis improved. As for an example, when lacking target sequence information, the polymerase chain reaction (PCR) based methods can greatly increase the feasibility of hig-throughput marker screening, relying on arbitrary primers. However, the genome of *B. distachyon* had been released to the public [6], thus the PCR-based methods, in turn, were overtaken by the widespread adoption of more robust microsatellite or simple sequence repeat (SSR) markers [7]. Furthermore, single nucleotide polymorphisms (SNPs) have more recently replaced SSR markers as genomic sequence data available [8–11]. Ongoing improvements in DNA sequencing have greatly accelerated sequence-based markers application and engage a continued convergence of sequencing and genotyping technologies. Under the second- and third-generation DNA sequencing technology improvement [12, 13], they promise delivery of technology for routine sequencing of genus *Brachypodium* genomes enabling evolution analysis within this genus.

Molecular markers were widely applied in evolutionary researches on plants, which undoubtedly became very useful tools in the identification of species and determining phylogenetic relationships. Under the determination of genetic relationships, evolutionary and conservational analyses of *Brachypodium* were also easy to be investigated. Furthermore, they could directly support the breeding of *Brachypodium* by marker-assisted selection. This chapter will outline the methods of molecular marker techniques and their applications to *Brachypodium*.

2 Materials

2.1 Samples Collection

In order to obtain quality genomic DNA, fresh and young leaves were obtained from fields or planted greenhouse. The materials should not be clean and without infection. All the samples should be properly labeled, placed in polybags, and immediately air-dried in silicon for short-term storage. For long-term storage, they should be ideally lyophilized at liquid N_2 and stored at −20 °C.

For the genus *Brachypodium*, different species were easy to be misidentified due to similar morphology. Thus, voucher specimens should be identified by professional taxonomists before dwelling into further studies.

For the below-mentioned study, leaf samples of fully grown plants of *B. sylvaticum* var. *sylvaticum* were collected from natural populations growing in the northwest of Yunnan Province, P. R. China. Identification was done on the basis of morphological characters of leaf and fruit by the taxonomists of KIB (Kunming Institute of Botany, KIB, CAS). A minimum distance of 10 m should be addressed in sample collection to avoid the colony.

2.2 Genomic DNA Isolation and Quantification (See Notes 1 and 2)

1. 2× genomic DNA extraction buffer: (2% cetyltrimethyl ammonium bromide (CTAB) (Sigma Chemical Company, St. Louis, USA), 100 mM Tris–HCl, pH 8, 20 mM ethylene diamine tetraacetic acid (EDTA), pH 8.0, 1.4 M NaCl).
2. Chloroform:isoamyl alcohol (24:1).
3. 100% ethanol or isopropanol.
4. 70% alcohol.
5. 10 mM TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8.0).
6. RNase A (10 mg/mL) (New England Biolabs Inc., MA, USA).
7. 50× Tris–Acetate-EDTA (TAE) buffer (pH 8.0).
8. Agarose (Promega Corporation, MI, USA).
9. Ethidium bromide (10 mg/mL).
10. 6× loading dye (30% glycerol, 5 mM EDTA, 0.15% bromophenol blue, 0.15% xylene cyanol).
11. λ/*Hin*dIII DNA ladder (New England Biolabs Inc., MA, USA).

2.3 Simple Sequence Repeat (SSR)

1. Genomic DNA Isolation and Quantification (*see* Subheading 2.2).
2. Distilled and Milli-Q water.
3. *Taq* DNA polymerase with 10× buffer supplied with 25 mM $MgCl_2$ (available from several suppliers).
4. 10 mM dNTPs: 10 mM each of dATP, dCTP, dGTP, and dTTP (Promega Corporation, MI, USA)
5. 10 μM Primers: Dilute in water or TE buffer to a 10 μM concentration.
6. Bench microcentrifuge.
7. Programmable thermal cycler supplied with heat cover.

2.4 Amplified Fragment Length Polymorphism (AFLP)

1. Genomic DNA Isolation and Quantification (*see* Subheading 2.2).
2. Distilled and Milli-Q water.

3. Two restriction enzymes: One rare (6 bp) and one frequent (4 bp) cutter and appropriate reaction buffer (available from several suppliers) (*see* **Note 3**).
4. Adaptor/ligation solution (0.4 mM ATP, 10 mM Tris–HCl, pH 7.5, 10 mM Mg-acetate, 50 mM K-acetate, dissolved in final volume of 10 mL with water).
5. dNTP mix (10 mM each, available from several commercial suppliers with optimal reaction buffer); T4 DNA ligase (1 U/μL).
6. Matching adaptors: *Eco*RI/*Mse*I adaptors at 5 and 50 pmol/μL concentration, respectively (Reactions were performed according to the instruction of the commercial enzymes).
7. AFLP-selective pre-amplification primer mixture (*Eco*RI pre-forward primer,

   5′-GACTGCGTACCAATTC**A**-3′; *Mse*I pre-reverse primer, 5′- GATGAGTCCTGAGTAA**C** -3′; underlined, enzyme-specific region; bold, selective nucleotide): Dilute each primer in water or TE buffer to a 50 ng/μL concentration.
8. AFLP-selective amplification primer mixture (*Eco*RI forward primer,5′- GACTGCGTACCAATTCA**NN**-3′; *Mse*I reverse primer,5′-GATGAGTCCTGAG TAA**CNN** -3′; underlined, enzyme-specific region; bold, selective nucleotide): Dilute each primer in water or TE buffer to a 30 ng/μL concentration.
9. *Taq* DNA polymerase (5 U/μL) and appropriate reaction buffer (available from many commercial suppliers) (*see* **Note 4**).
10. 1× TAE buffer; 500 mM EDTA (pH 8.0); 1 M Tris–HCl (pH 8.0); TE buffer: 10 mM Tris–HCl, pH 7.5, 1 mM EDTA. To make 100 mL of TE buffer, add 1 mL 1 M Tris–HCl pH 7.5 and 200 μL 500 mM EDTA pH 8.0 to 98.8 mL water, mix, and autoclave before use.
11. Bench microcentrifuge (Eppendorf, Hamburg, Germany).
12. Programmable dry incubators or water baths (Eppendorf, Hamburg, Germany).
13. Programmable thermal cycler supplied with heat cover (Eppendorf, Hamburg, Germany).
14. Microwave oven (Eppendorf, Hamburg, Germany).
15. Standard horizontal agarose gel electrophoresis apparatus (Eppendorf, Hamburg, Germany).
16. Power supply (Eppendorf, Hamburg, Germany).

2.5 Inter-Simple Sequence Repeats (ISSR)

1. Genomic DNA Isolation and Quantification (*see* Subheading 2.2).
2. MilliQ water.

3. dNTP mix (10 mM each): Add 10 μL of each 100 mM dNTP solution to 60 μL water.
4. *Taq* DNA polymerase (5 U/μL) with 10× reaction buffer (available from many commercial suppliers).
5. ISSR primer: Dilute in water or TE buffer to a 20 μM concentration (*see* **Note 5**).
6. Bench microcentrifuge (Eppendorf, Hamburg, Germany).
7. Programmable thermal cycler supplied with heat cover (Eppendorf, Hamburg, Germany).

2.6 Electrophoresis

1. 40% Acrylamide:bis-acrylamide (29:1).
2. 7.5 M Urea.
3. 10× Tris-Borate-EDTA (TBE) buffer, pH 8.0.

2.7 PAGE Reagents

1. 40% Acrylamide bis-acrylamide.
2. 7.5 M Urea.
3. 10× Tris–Borate-EDTA (TBE) buffer, pH 8.0.
4. Cover the bottle with aluminum foil and store at 4 °C and use before 1 month.
5. 100 bp DNA ladder (New England Biolabs Inc., MA, USA).

2.8 Silver Staining Reagents

1. Acetic acid, glacial.
2. Silver nitrate crystal, AR (ACS) ($AgNO_3$).
3. Formaldehyde solution, AR (ACS) (HCHO).
4. Sodium thiosulfate (Na_2S_2O).
5. Sodium carbonate powder, ACS reagent (Na_2CO_3).
6. Ethanol.
7. Silver staining solution (250 mg silver nitrate and 375 μL formaldehyde and 50 μL sodium thiosulfate).
8. Ice-cold developer solution (10 °C) (7.5 g sodium carbonate, 375 μL formaldehyde, and 50 μL sodium thiosulfate (10 mg in 1 mL water) in 250 mL water).
9. Formamide loading dye (80% formamide, 10 mM EDTA pH 8.0, 1 mg/mL Xylene cyanol 1 mg/mL, bromophenol blue 50 mg).

3 Methods

3.1 Isolation of Genomic DNA

The genomic DNA extraction of genus Brachypodium is according to the CTAB DNA isolation protocol with slightly modified for optimum yield [14].

1. Lyophilized leaves (200 mg) are grounded to a fine powder in liquid N_2 using mortar and pestle.
2. Transfer the leaf powder to sterile polypropylene tubes containing 20 mL of pre-warmed (65 °C) CTAB extraction buffer and mix gently but thoroughly till no clump was visible.
3. Incubate for 40 min at 65 °C in a water bath with regular stirring every 5 min.
4. Add an equal volume of chloroform and mix thoroughly.
5. Centrifuge at 4300 × *g*, at 20 °C for 20 min.
6. Measure the upper aqueous phase and transfer to a sterile polypropylene tube.
7. Add RNase A (10 mg/mL) (New England Biolabs Inc., MA, USA) to a final concentration of 100 μg/mL and incubate at 37 °C for 30 min.
8. Re-extract the samples with an equal volume of chloroform and centrifuge at 2420 × *g*, at 20 °C for 20 min.
9. Transfer the aqueous phase to a sterile polypropylene tube and measure its volume.
10. Precipitate the DNA by adding an equal volume of ice-cold isopropanol to the aqueous phase and incubate at −20 °C for 30 min.
11. Precipitate the DNA centrifuged at 6700 × *g*, at 4 °C for 30 min.
12. Discard the supernatant and wash the DNA pellet with 70% ethanol.
13. Air dry the DNA pellet and dissolve in 200-μL sterile double distilled water.
14. Store at −20 °C till further use.

3.2 DNA Qualification

It is an essential step for following performing PCR techniques with known amount of DNA (*see* **Note 6**).

The comparison of an aliquot of the extracted sample with standard DNAs of known concentration (λ/*Hin*d III) (New England Biolabs Inc., MA, USA) can be done using gel electrophoresis.

1. 5 μL of the DNA is mixed with 1 μL of 6× loading dye and loaded onto a 0.8–1% agarose gel along with 500 ng of λ/*Hin*d III marker (New England Biolabs Inc., MA, USA) and electrophoresis at 90 V for 30 min.
2. The quantity of extracted DNA is estimated based on the intensity of λ/*Hin*d III digest marker bands as the top bands account half amount (250 ng) of the total loaded amount.
3. The quality of genomic DNA is confirmed for its integrity.

3.3 Primers Design

Primers can be obtained commercially as custom oligonucleotides (*see* **Note 7**) and should be designed from available genomic DNA sequence to fulfill the following set of criteria:

1. Forward primer and reverse primer flank the SSR sequence.
2. Primer length 18–25 bp.
3. GC content of primer >40%.
4. Annealing temperature of primer >45 °C.
5. No strings of repeated mononucleotides >3.
6. No repetitive regions or regions which when inverted will bind to each other.
7. No complementary sequences between the forward and reverse primers.

3.4 Sequence Amplification for SSR (See Note 8)

1. Make up 10–50 μL (*see* **Note 9**) per reaction PCR mixes in 0.2 mL PCR tubes on ice containing (*see* **Note 10**):
   (a) 0.125 mM dNTP mix (e.g., 1.6 μL of 2.5 mM solution in a 20 μL reaction).
   (b) 1× DNA polymerase buffer premix with 20 mM $MgCl_2$ (e.g., 2 μL of 10× DNA polymerase buffer in a 20 μL reaction).
   (c) 0.5 μM forward primer (e.g., 1 μL of 10 μM solution in a 20 μL reaction).
   (d) 0.5 μM reverse primer (e.g., 1 μL of 10 μM solution in a 20 μL reaction).
   (e) 10–75 ng of genomic DNA (e.g., 5 μL of 10 ng/μL DNA solution in a 20 μL reaction).
   (f) 1 U of *Taq* (DNA polymerase enzyme from *Thermus aquaticus*; e.g., 0.2 μL of 5 U/μL solution in a 20 μL reaction) (New England Biolabs Inc., MA, USA).
   (g) Purified deionized water to the appropriate volume (e.g., up to 20 μL in a 20 μL reaction).
2. Mix thoroughly by flicking the tube (do not vortex).
3. Thermocycler programming (*see* **Note 11**).

Heat cycling (*see* **Note 12**): Initial denaturation, then 15–35 cycles of denaturation, melting, and annealing, followed by a final extension. Typical temperatures and times are given for a product of 300 bp, and the melting time will change by the primers design.

(a) Initial denaturation: 94 °C for 5 min. Then 35 cycles of **steps 2–4**.
(b) Denaturation: 94 °C for 30 s.

(c) Melting (*see* **Note 13**): 50 °C for 45 s.

(d) Annealing: 72 °C for 60 s.

Followed by a single extension step:

(e) Extension: 72 °C for 10 min.

(f) Stop: store at 4 °C.

3.5 Amplified Fragment Length Polymorphism

1. DNA digest and adaptor ligation were according to the description by Vos et al. [15].
2. To prepare the template for selective AFLP amplification, combine, for each reaction product, 3 μL of the pre-selective amplification reaction product with 147 μL 1 mM TE buffer.
3. Mix by gently tapping the tube and spin for 10 s.
4. Store at 2–6 °C until use (*see* **Note 14**).
5. For each reaction, add 5 μL of the diluted pre-selective PCR products template to a 0.5 or 0.2 mL PCR microtube.
6. Add 2 μL of 10× PCR buffer.
7. Add 1 μL of each selective primer (*Mse*I reverse + *Eco*RI forward primer).
8. Add 0.4 μL 10 mM dNTPs mix.
9. Add 2 μL 1.5 mM $MgCl_2$ (*see* **Note 15**).
10. Add 0.2 μL (1 U) *Taq* DNA polymerase.
11. Mix gently and centrifuge briefly to collect reactions at the bottom of the tube.
12. Complete to 20 μL with water.
13. Amplify by means of a touch-down PCR as follows: one cycle of 94 °C denaturation for 30 s, 65 °C annealing for 30 s, and 72 °C extension for 60 s, followed by 12 cycles with the annealing temperature lowered by 0.7 °C per cycle. Complete with 23 further cycles of 94 °C for 30 s, 58 °C for 30 s, and 72 °C for 60 s (*see* **Notes 16** and **17**). Set soak temperature to 12 °C (*see* **Note 18**). Keep the samples in the fridge until electrophoresis.

3.6 Inter-Simple Sequence Repeats [16]

1. For a standard reaction (20 μL volume), add 2 μL of 10× PCR buffer with $MgCl_2$ supplied to a 0.5 or 0.2 mL PCR microtube.
2. Add 1 μL of each primer (*see* **Note 19**).
3. Add 0.4 μL 10 mM dNTPs mix.
4. Add 0.2 μL (1 U) *Taq* DNA polymerase.
5. Complete to 18 μL with water.
6. Mix by gently tapping the tube and spin briefly to collect the mix in the bottom of the tube.
7. Aliquot 18 μL to each tube and add 2 μL of respective DNA sample (diluted at 10–20 ng/μL).

3.7 Preparation and Casting of Polyacrylamide Gels (See Note 20)

It is necessary to avoid skin contact with acrylamide for it is cancerigenous and neurotoxic, thus, while handling these products, protective gloves and eyewear should be worn. Carry out all procedures at room temperature.

1. Prepare the denaturant polyacrylamide gel solution: Prepare 60 mL of gel (6% acrylamide, 7.5 M urea, 1× TBE, 0.7% APS, 0.05% TEMED) combining 45.8 mL of 9.6 M urea (*see* **Note 21**) with 6 mL 10× TBE and 7.2 mL acrylamide solution 50% in a 100 mL beaker, mixing by gentle stirring (*see* **Note 22**). Add 450 μL of 10% ammonium persulfate (APS) to the mixture followed by 32 μL of TEMED. Mix quickly but gently (*see* **Notes 23** and **24**). Immediately, cast the prepared polyacrylamide gel using a sequencing apparatus (*see* **Note 25**).
2. Using a sterile plastic pipette, carefully pour the gel solution between the assembled glass plates. Keep the assembled plates with ca 45 °C angle in relation to both vertical and horizontal planes and always pour the solution at one side with a constant flow to prevent bubble formation (*see* **Note 26**). If any bubbles are noticeable, gently tap the glass plate or move the assembled cast to remove them.
3. Once the cast is filled up with the gel solution, insert the comb (s) into the gel with the teeth facing up (*see* **Note 27**). Start inserting the comb(s) by the edge of the plate. Clamp with three clips and keep it at a 5° angle relative to the surface while the gel polymerizes (*see* **Notes 28** and **29**).
4. After the acrylamide has polymerized, remove the clamps holding the comb(s) and casting stand.
5. Pull out the comb(s) straight by wriggling it gently and smoothly.
6. Remove the glass holding clips or casting clamp.
7. Place the stand with the smaller glass facing back, into the apparatus tank, touching the base of the lower buffer reservoir.
8. Fill the upper and lower reservoirs with 1× TBE buffer (*see* **Note 30**).
9. Mark the level of the buffer in the upper chamber with a pen marker for subsequent check for possible leakage (*see* **Note 31**).
10. Gently flush the wells thoroughly with running buffer using a discardable Pasteur pipette (*see* **Note 32**).
11. Gently insert the shark toothcomb between the glass plates with teeth facing downwards.
12. Fix the safety cover on top on the upper buffer chamber to prevent evaporation of buffer and pre-run the gel at constant power (e.g., ca. 55 W) for ca. 20 min.

13. For electrophoresis under denaturing conditions (to resolve AFLP and SAMPL products), denature the PCR products (5 μL each sample); prepare two tubes with molecular weight marker (1 μL of 100 bp ladder) mixed with equal volume of 2× denaturant loading dye for 3 min at 95 °C. Immediately, transfer the denatured samples to ice to prevent annealing.
14. Load the samples into each well. Load 3–5 μL of SSR/AFLP/ISSR products. Load the two extreme wells with the appropriate DNA ladder.
15. After loading all the samples and markers, close the lid of the upper buffer chamber.
16. Allow the gel to run at constant power until the xylene cyanol-dye reaches ca. 2/3 of the gel (*see* **Notes 33–35**).
17. Remove the plates carefully from apparatus and remove excess buffer by blotting the bottom of the cast on a stack of paper towels.
18. Remove the spacers and separate the plates carefully so that the gel should retain on the smaller glass plate (*see* **Note 36**).
19. Take the larger glass plate with the gel attached and proceed to the gel silver staining protocol to visualize the bands.

3.8 Silver Staining of Gels

As silver staining is a temperature-dependent process, lab temperature should be controlled (to 18–24 °C). Keep all silver staining solutions protected from light. Carry out all procedures, including gel agitation, in a fume hood and wear gloves at all stages. The solution recipes are given per 2 L to be directly suited for staining gels of the dimensions used.

1. Before beginning, put the fixing/stopper, the impregnation, and the developer solutions at 4 °C (*see* **Note 37**). Keep an additional 1 L of water refrigerated at 4 °C.
2. Fix the gel with 2 L of fixing solution during at least 30 min (*see* **Notes 38** and **39**).
3. Rinse with 2 L deionized water three times × 2 min.
4. Incubate gel for 20 min in 4 °C cold silver impregnation solution for 30 min. Keep the gel in the same tray during all subsequent steps.
5. Rinse once with dH_2O for no more than 15 s (*see* **Note 40**).
6. Develop by soaking the gel with developer solution until the bands are revealed (typically 2–5 min; *see* **Note 41**).
7. Stop the reaction by dispersing 1 L of 10% acetic acid in the developer solution (stopper solution, reserved from **step 2**) for at least 5 min.

8. Wash with deionized water for 10 min and let dry on the bench in a near vertical position.
9. Digitally scan the gel for a permanent record (*see* **Note 42**).

3.9 Analyzing Plant SSR Data

1. Sequence the individuals of all populations to get the DNA lengths of SSR alleles.
2. Collate sequence length data for all SSR markers run on the population, and sort by sample collection location.
3. Data may now be inputted into one of a range of free or commercial software programs in order to perform analyses such as creation of linkage maps/determination of population linkage disequilibrium, determination of population genetic diversity, creation of phylogenic relationship trees, and population principal components analysis (PCA). The addition of phenotype or population structure data (e.g., cultivars, families) may allow analyses, association mapping, matching of haplotypes to known phenotypes, and correlation analyses (*see* **Note 43**).

   Example: Genetic Diversity Analysis by using SSR Markers on *Brachypodium sylvaticum* var. *breviglume* from Yunnan, China. *B. sylvaticum* var. *breviglume* is a perennial self-compatible diploid bunchgrass which widely distributed in Southwestern China. These species exhibit divergently ecological adaptation, frequently intraspecific hybridization, and high levels of phenotypic variation. In this study, we aim to identify and characterize new microsatellite markers for the characterization of genetic diversity and population structure of *B. sylvaticum* var. *breviglume* to obtain a better knowledge of adaptation, evolution, and diversification of the species in Southwestern China. The materials and methods were turned to be described by Mo et al. (2013) [17]. The microsatellites which were successfully amplified and generated clear fragments were further used to detect polymorphisms for all the individuals from the five natural populations. They were screened on 8% denaturing polyacrylamide gels and visualized by silver staining with a 100 bp extended DNA ladder as a size standard (Fig.1).

   Genetic diversity of SSR loci including percentage of polymorphic loci (*PPL*), observed heterozygosity (H_O), expected heterozygosity (*He*), Hardy–Weinberg equilibrium, Wright's F_{ST}, and the Wright's fixation index (F_{is}) were calculated according to Mo et al. (2013). Factorial correspondence analysis (FCA) was also estimated.

   The result showed that the high levels of polymorphism were detected by the SSR loci detection, with the total number of alleles equaling 94 (Table 1). The expected heterozygosity within populations (mean $He = 0.488$) was higher than that of *B. sylvaticum* (Huds.) Beauv. from Oregon (mean

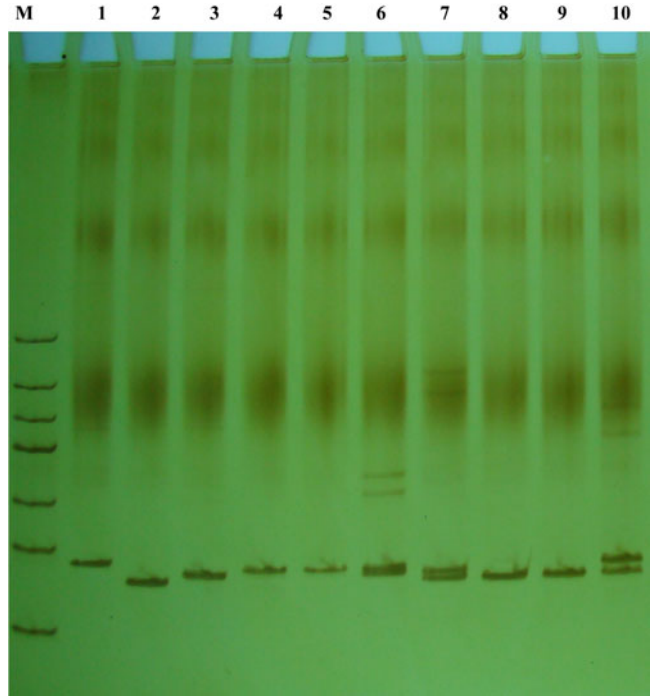

Fig. 1 The PAGE analysis of PCR products by using different SSR loci
M: 100 bp DNA ladder 1–10: PCR products of different individuals of *B. sylvaticum* var. *breviglume* Keng. Note: One band revealed that the individual was homozygote and two bands showed that of heterozygote.

He = 0.349) [18] (Table 2). In general, the observed low *F*st within the study populations and high allelic diversity and heterozygosity estimates suggest that genetic drift has not yet had a major influence on the study populations. The amount of genetic diversity present among the population was considerable (*F*st = 0.488) higher than the values detected in natural populations of *B. sylvaticum* (Huds.) Beauv. from Oregon (*F*st = 0.441) and lower than that of invasive populations (*F*st = 0.493). The *F*is value (0.784) detected here in *B. sylvaticum* var. *breviglume* Keng is higher than that observed in native (*F*is = 0.447) and invasive populations (*F*is = 0.616) of *B. sylvaticum* (Huds.) Beauv. from Oregon, which indicated that *B. sylvaticum* var. *breviglume* Keng was a highly self-compatible species.

4 Notes

1. Successful molecular marker technique is relied on the complete restriction digestion of genomic DNA; thus, high quality and intact genomic DNA should be taken seriously, without

Table 1
Ecological and geographical parameters and levels of genetic variation of the five studied populations of *B. sylvaticum* var. *breviglume* across the seven SSR loci

Population codes	Location	Latitude (N)	Longitude (E)	*N*	*Na*	*Ne*	*PPL*	*Ho*	*He*
TH	Tanhua Mountain, Dayao County, Yunnan Province	25°57′	101°13′	20	2.86	2.03	85.71	0.143	0.439
YY	Yangyu Mountain, Eryuan County, Yunnan Province	26°10′	100°18′	20	3.43	2.53	100	0.220	0.520
ML	Meili Mountain, Deqin County, Yunnan Province	28°24’	98°45’	20	3.29	2.37	100	0.090	0.561
JZ	Jizu Mountain, Binchuang County, Yunnan Province	25°57’	100°23’	20	3.29	2.40	100	0.044	0.441
YL	Yulong Mountain, Lijiang City, Yunnan Province	27°00’	100°11’	20	2.86	2.16	100	0.029	0.481
Average				20	3.14	2.30	97.14	0.105	0.488

N sample size, *Na* mean number of alleles, *Ne* mean number of effective alleles, *PPL* percentage of polymorphic loci, *Ho* observed heterozygosity, *He* expected heterozygosity

Table 2
Measurements of genetic diversity, genetic differentiation, and gene flow among populations of *B. sylvaticum* var. *breviglume* across the seven microsatellite loci

Locus	*Na*	*Ne*	H_o	H_e	F_{is}	F_{st}	N_m
2-3A1	5.600	3.971	0.128	0.728	0.824	0.136	1.586
2-6E6	3.400	2.129	0.056	0.486	0.885	0.258	0.718
2-6E8	3.800	2.622	0.217	0.584	0.628	0.352	0.459
3-2B2	2.400	2.284	0.092	0.522	0.824	0.328	0.513
3-2E3	1.800	1.325	0.052	0.206	0.750	0.671	0.123
BS537	2.800	2.016	0.121	0.499	0.758	0.266	0.691
BS545	2.200	1.736	0.071	0.392	0.820	0.306	0.567
Mean	3.14	2.30	0.105	0.488	0.784	0.331	0.665

Note: F_{is} deficiencies of heterozygotes relative to Hardy–Weinberg expectations, *Nm* gene flow estimated from *Fst* ($Nm = 0.25 \times (1-Fst)/Fst$)

contaminating nucleases or inhibitors. DNA should be dissolved in ultrapure MilliQ water or TE pH 8.0 buffer.

2. All the solutions preparation were according to methods described on sector of reagents and solutions in the book "Molecular Cloning" [19]. All the reagents were provided by local reagent supply company which is mentioned in this chapter except the special labeled reagents.
3. Selecting a 4 bp- and 6 bp-generated restriction enzymes produces small DNA fragments in the optimal size range (50 bp–1 kb) to be amplified and easily separated on denaturing polyacrylamide gels. For typical AFLP procedures, *Eco*RI and *Mse*I enzymes are common used. Other enzymes can be selected but, preferably, they should not be methylation-sensitive. Make sure that the chosen enzymes are active in the same reaction buffer.
4. Most of the current *Taq* DNA polymerases in the market can be used, however, some of them will fail to produce satisfactory results and need to be tested before large screening.
5. Di-, tri-, or tetra-repeats can be used to design primers. Anchored primers are recommended to avoid floating of the primer in the satellite sequence. Primers anchored in the 3′ end will produce polymorphisms considering both the length of the satellite and the distance between satellites. Typical ISSR primers are (5′–3′) as follows: (GA) 8 YG, (AG) 8 YC, (AG) 8 YT, (CA) 8 R, (AGC) 4 YT, (AGC) 4 YR, VHV(GT) 7, VHV(TG) 7,HVH(CA) 7, HVH(TG) 7, DBD(AC) 7, and (AGC) 4 YR (Y = pyrimidine, B = every base except A, D = every base except C, H = every base except G, V = every base except T).
6. The most important factor for reproducibility of the SSR profile has been found to be the result of inadequately prepared template DNA which could be overcome through choice of an appropriate DNA extraction protocol to remove any contaminants [20]. Differences between the template DNA concentrations of two DNA samples will result in the loss or gain of some bands.
7. Primers can be designed and ordered as custom oligonucleotides from those companies that provide these commercially. They often provide the product information that give an estimate of primer secondary structures, melting temperatures, and compatibilities between primer pairs based on primer sequences. Primers can be shipped at room temperature, and ordered at custom concentrations or volumes or as dried down pellets.
8. PCR failure will often happen in inexperienced operators and particularly in teaching labs is most commonly due to mistakes in master mix composition, such as failing to add a reagent or

adding the wrong volume of a reagent. It is essential for success PCR reaction that keeping the reagent mix cold and mixing well (flicking the tube is preferable to inverting, but vortexing is not acceptable) after addition of all components. Poor DNA quality (contaminants) is also common and lead to PCR failure, although generally the process is extremely tolerant of DNA quantity (1–200 ng will still work in many instances).

9. The minimum recommended PCR volume is usually 10 μL, and 50 μL is more than sufficient for molecular marker genotyping purposes. Lower reaction volumes are more likely to fail, and 20–25 μL reaction volumes provide a good compromise between success rates and savings on reagent costs.
10. For more than a few samples, make up a master mix containing all reagents except the genomic DNA. For example, when preparing 24 samples of DNA for PCR, multiply the amounts of all reagents except DNA required for one reaction volume by 25 (±1 for pipetting error), then add the appropriate amount of master mix (e.g., 15 μL in a 20 μL reaction volume containing 5 μL DNA solution) to individual DNA samples in PCR tubes.
11. The PCR mix should be repeatedly heated and cooled in cycles. Although this can be achieved by manually transferring tubes between appropriately heated water baths, commercially produced thermocyclers, which contain tube-holding blocks that can be heated and cooled to precise temperatures, are more commonly used.
12. Most commercial *Taq* DNA polymerase instructions suggest thermocycler protocols optimized for those particular enzymes, as well as tubes of 10× reaction buffer, $MgCl_2$ solution, and dNTP mix. A number of other modified *Taq* enzymes with "hot start" and "proof reading" capabilities are also commercially available. Hot start *Taq* is more thermostable and needs to be run at higher temperatures (e.g., 98 °C) for the denaturation and annealing steps of the thermocycler protocol. Proof-reading *Taq* makes less sequence errors during replication than regular *Taq*. Neither hot-start nor proof-reading *Taq* is required for genotyping procedures, although proof-reading *Taq* may provide a more robust means of checking polymorphisms during sequencing of isolated microsatellite regions in the primer design validation phase.
13. Melting temperature (T_m) is determined by the composition of the oligonucleotide primers. Longer primers with higher GC content will have higher melting temperatures compared to shorter primers with lower GC content, and the sequence will also affect the T_m. Commercial primers synthesis report will offer the information of the melting temperatures (T_m), and

compatibilities between primer pairs based on primer sequences, and for each pair of primers using the lower melting value is generally recommended. Lowering the melting temperature during the PCR will reduce primer binding specificity, and is hence more likely to produce a product in recalcitrant reactions and also to produce multiple, a specific products (especially in polyploids). Increasing the melting temperature will increase primer binding specificity, but the reaction may fail if the melting temperature is too high relative to the primer T_m.

14. Store the unused portion as aliquots, at −20 °C.
15. Commonly it can be omitted if the reaction buffer already contains magnesium.
16. Minor optimizations may be needed according to different polymerase brands and thermal cycler machines.
17. PCR is started at a relative high annealing temperature to obtain optimal primer selectivity. In the following steps, the annealing temperature is lowered gradually to a temperature at which efficient primer binding occurs. This temperature is then maintained for the rest of the PCR cycles.
18. Usually, the soaking temperature is 4 °C, but keeping the thermocycler at 12 °C saves energy and is not detrimental to the samples.
19. The conditions may require optimization. It should be noted that ISSR primers contain repetitive regions and can be more difficult to amplify, so increased concentrations of primer (up to 20 μM) may be necessary.
20. During the initial SSR primers screening, PCR product electrophoresis should be run on the polyacrylamide gels, and stained by the silver to visualize the band, and hence identified the primers pair suitability for following analysis.
21. Urea easily precipitates at low temperatures. Before use, check for total dissolution and, if needed, warm at 37 °C for about 30 min (or until dissolved).
22. Mix gently to prevent gas bubble formation. No degas is needed in our hands. If proved necessary, perform this step via vacuum application for 10 min, using a Kitasato flask.
23. Add the TEMED immediately prior to pouring the gel and work quickly after its addition to complete pouring the gel before the acrylamide polymerizes.
24. Most gel formulations allow only approximately 5 min before starting polymerization. Work quickly to be able to load all the gel solution inside the gel cast.

25. When pouring the gel from the top, make sure the bottom fill port of the casting clamp is sealed to prevent the gel solution to leak from the bottom.
26. Gently lower a bit of the angle of the glass plates while pouring the gel during casting.
27. This step aims at creating a perfectly flat surface in the top of the gel, permitting the sample to be uniformly loaded. Disturbances in the surface of the gel will result in "waving" of the bands.
28. Takes 2 h to overnight. Better results are obtained with an overnight polymerization. In such conditions, the gel should to be formed and immersed with 1× TBE (running buffer).
29. A simple way to monitor polymerization is to check for a small amount that was left not casted. Typically, the acrylamide polymerizes completely for 45–120 min at room temperature.
30. The volume of running buffer should be enough to cover the combs inserted to generate the flat surface.
31. Make sure that the glass plates are firmly seated against the inner surface of the casting clamp. Verify that the upper buffer chamber drain valve is in the closed position (down), and fill the upper buffer chamber with running buffer, covering the gel. Make sure that no leakage exists from the upper buffer chamber. Do not start electrophoresis if leakage is observed. If leakage is observed during the gel run, keep adding running buffer to the top chamber of the apparatus to prevent overheating and cracking of the glasses or, if it is not possible, abort electrophoresis immediately.
32. Use a 200 μL micropipette, a discardable Pasteur pipette or a syringe filled with 1× TBE buffer with an attached needle to flush out all the wells. Pipette vigorously up and down several times. Acrylamide gel fragments will prevent homogenous run, leading to distortion of the bands.
33. At 45 W and under our conditions, a complete run takes about 2–2.30 h for SSR and AFLP and 2.30–3 h for ISSR.
34. During the electrophoresis, keep monitoring possible buffer leakage and gel temperature. An appropriate indicator placed onto the outer plate near the center of the gel can be used as an option. The temperature should be maintained between 40 and 50 °C.
35. Xylene cyanol co-migrates with ca. 125 bp linear single-stranded DNA of in 6% denaturing gels and co-migrates with ca. 230 bp linear double-stranded DNA in 6% non-denaturing gels.

36. The best way to separate the glass plates is to use a pizza wheel. Carefully insert the wheel between the two glasses (use the spacers to create enough room for the wheel at the top contact between the two glass plates) and gently and smoothly wriggle it until the plates separate.
37. Staining is enhanced with cold $AgNO_3$.
38. Fix until no dye is visible on the gel.
39. After fixation, reserve 1 L of the solution to be used in **step 7**.
40. Residual $AgNO_3$ on the gel surface and staining tray will increase background staining.
41. The developing time varies. Larger bands (top of the gel) develop first. A certain (but controlled) degree of over staining in this area of the gel is acceptable in order to visualize the smaller bands.
42. The silver staining for SSR was performed during the SSR primers screened, but for AFLP and ISSR, the data should be recorded and analyzed for further identification.
43. Logistic rather than normal linear regression should be performed using SSR data, as SSR alleles are binomially rather than normally distributed. Data cleaning to remove alleles with a high degree of failed amplification across the population is also suggested, as these may bias subsequent analyses.

References

1. Henry RJ, Edwards M, Waters DL, Gopala Krishnan S, Bundock P, Sexton TR, Masouleh AK, Nock CJ, Pattemore J (2012) Application of large-scale sequencing to marker discovery in plants. J Biosci 37(5):829–841
2. King RC, Mulligan P, Stansfield W (2013) A dictionary of genetics. Oxford University Press, New York
3. Semagn K, Bjørnstad Å, Ndjiondjop MN (2006) An overview of molecular marker methods for plants. Afr J Biotechnol 5 (25):2540–2568
4. Nielsen G (1985) The use of isozymes as probes to identify and label plant varieties and cultivars. Isozymes Curr Top Biol Med Res 12:1–32
5. Brown AH (1978) Isozymes, plant population genetic structure and genetic conservation. Theor Appl Genet 52(4):145–157. doi:10.1007/BF00282571
6. International Brachypodium I (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463 (7282):763–768. doi:10.1038/nature08747
7. Garvin DF, McKenzie N, Vogel JP, Mockler TC, Blankenheim ZJ, Wright J, Cheema JJ, Dicks J, Huo N, Hayden DM, Gu Y, Tobias C, Chang JH, Chu A, Trick M, Michael TP, Bevan MW, Snape JW (2010) An SSR-based genetic linkage map of the model grass Brachypodium distachyon. Genome 53(1):1–13. doi:10.1139/g09-079
8. Ma JQ, Huang L, Ma CL, Jin JQ, Li CF, Wang RK, Zheng HK, Yao MZ, Chen L (2015) Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified Fragment sequencing (SLAF-seq). PLoS One 10(6): e0128798. doi:10.1371/journal.pone.0128798
9. Rellstab C, Zoller S, Tedder A, Gugerli F, Fischer MC (2013) Validation of SNP allele frequencies determined by pooled next-generation sequencing in natural populations of a non-model plant species. PLoS One 8(11): e80422. doi:10.1371/journal.pone.0080422
10. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their

impact on plant breeding. Int J Plant Genomics 2012:728398. doi:10.1155/2012/728398

11. Henry R, Edwards K (2009) New tools for single nucleotide polymorphism (SNP) discovery and analysis accelerating plant biotechnology. Plant Biotechnol J 7(4):311. doi:10.1111/j.1467-7652.2009.00417.x
12. Ueno S, Moriguchi Y, Uchiyama K, Ujino-Ihara T, Futamura N, Sakurai T, Shinohara K, Tsumura Y (2012) A second generation framework for the analysis of microsatellites in expressed sequence tags and the development of EST-SSR markers for a conifer, Cryptomeria japonica. BMC Genomics 13:136. doi:10.1186/1471-2164-13-136
13. Solignac M, Mougel F, Vautrin D, Monnerot M, Cornuet JM (2007) A third-generation microsatellite-based linkage map of the honey bee, Apis mellifera, and its comparison with the sequence-based physical map. Genome Biol 8 (4):R66. doi:10.1186/gb-2007-8-4-r66
14. Springer NM (2010) Isolation of plant DNA for PCR and genotyping using organic extraction and CTAB. Cold Spring Harb Protoc 2010(11):pdb prot5515. doi:10.1101/pdb.prot5515
15. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23 (21):4407–4414
16. Zietkiewicz E, Rafalski A, Labuda D (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20(2):176–183. doi:10.1006/geno.1994.1151
17. Mo X, Gao J, Gao L (2013) Characterization of microsatellite markers and their application to genetic diversity analysis of Brachypodium sylvaticum Var. Breviglume from Yunnan, China. Am J Plant Sci 04(7):1427–1434
18. Rosenthal DM, Ramakrishnan AP, Cruzan MB (2008) Evidence for multiple sources of invasion and intraspecific hybridization in Brachypodium sylvaticum (Hudson) Beauv. In North America. Mol Ecol 17(21):4657–4669. doi:10.1111/j.1365-294X.2008.03844.x
19. Green MR, Sambrook J (2012) Molecular cloning: a laboratory manual, 4th edn. Cold Spring Harbor Laboratory Press, New York
20. Li H, Li J, Cong XH, Duan YB, Li L, Wei PC, XZ L, Yang JB (2013) A high-throughput, high-quality plant genomic DNA extraction protocol. Genet Mol Res 12(4):4526–4539. doi:10.4238/2013.October.15.1

Chapter 11

Estimate Codon Usage Bias Using Codon Usage Analyzer (CUA)

Zhenguo Zhang and Gaurav Sablok

Abstract

One amino acid is added to a growing peptide by a ribosome through reading triple nucleotides, i.e., a codon, each time. Twenty species of amino acids are often coded by 61 codons, so one amino acid can be coded by more than one codon and the codons coding the same amino acid are called synonymous. Intriguingly, synonymous codons' usage is often uneven: some are used more often than their alternatives in a genome. The unevenness of codon usage is termed codon usage bias (CUB). CUB is widespread, and its causes and consequences have been under intensive investigation. To facilitate the studying of CUB, in this chapter we present a protocol of estimating CUB by using the free software Codon Usage Analyzer, and apply it to *Brachypodium distachyon* as an example. To accomplish this protocol, the readers need some basic command-line skills. Briefly, the protocol comprises four major steps: downloading data and software, setting up computing environment, preparing data, and estimating CUB.

Key words Codon usage bias, Codon usage analyzer, Synonymous codons

1 Introduction

Codon usage bias (CUB) has been observed in many species, from prokaryotes to eukaryotes [1, 2]. The evolutionary causes and the functional consequences of CUB have been extensively studied, but remain elusive [3–13]. Given that the number of sequenced genomes is rapidly growing, now CUB can be studied in numerous organisms. This protocol, using *Brachypodium distachyon* as an example, demonstrates the procedures of estimating CUB for all genes in a genome.

In fact, numerous CUB metrics have been developed; the popular ones are Codon Adaptation Index (*CAI*) [14], tRNA Adaptation Index (*tAI*) [15], Frequency of optimal codons (*Fop*) [16], and Effective Number of Codons (*ENC*) [17]. These metrics are based on different assumptions regarding codon optimality. For example, *CAI* assumes that the codons used more frequently in highly expressed genes are translationally more efficient than

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_11,

synonymous alternatives, while *tAI* assumes codons matching most abundant isoaccepting tRNAs are optimal. The calculation of ENC borrowed a thought from effective population size in population genetics and the value ranges from 20 (strongest bias) to 61 (even codon usage). However, as shown later in this protocol, these metrics are usually highly correlated, so most time using any of them should suffice, but sometimes using different metrics may lead to different conclusions, so it is always wise to confirm a conclusion with all the metrics. In this protocol, we demonstrate how to calculate the above four metrics using the software Codon Usage Analyzer.

2 Materials

To calculate all four CUB metrics in *Brachypodium distachyon*, we need to obtain the necessary data and software from difference sources, all of which are publicly available. All the data used in this protocol can also be downloaded from https://github.com/fortune9/CUB_Brachypodium_distachyon. The protocol has been fully tested in Linux/Unix and should work in Mac OS, too.

2.1 Collecting Data

1. Create a folder "cub", which is the root folder for this protocol. It can be achieved by running the following command in a terminal window:

   ```
   $ mkdir cub
   ```

   Then download and put the following data into this "cub" folder. The folder structure will eventually look like that in Fig. 1, containing both input and output data.
2. CDS sequences, genome sequence, and GFF-formatted exon annotation of *Brachypodium distachyon*: these data are downloadable from the Phytozome website at http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Bdistachyon (*see* **Note 1**). The annotation version used in this protocol is Phyotozome 10.3, Brachypodium distachyon v2.1. The files downloaded are Bdistachyon_283_v2.1.cds_primaryTrancsriptOnly.fa.gz, Bdistachyon_283_assembly_v2.0.softmasked.fa.gz, and Bdistachyon_283_v2.1.gene_exons.gff3.gz. Due to alternative splicing, one gene may have multiple CDS sequences. In this protocol, we use the CDS of the primary sequence (defined in the Phyotozome annotation) for each gene.
3. RNA-seq raw reads: the fastq file for estimating gene expression level is downloaded from the NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra/ with accession number SRR921317) (also *see* **Note 7**). The file contains sequenced reads from the total RNAs in the 12-day-old Brachypodium seedlings using Illumina HiSeq 2000.

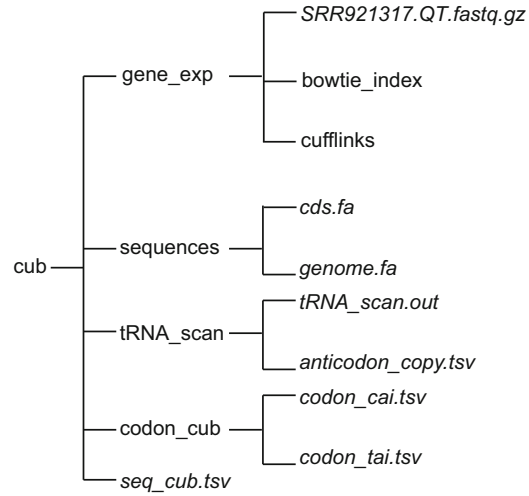

Fig. 1 The folder structure. The folders and plain files are presented in normal and italic fonts, respectively. Some plain files such as those in bowtie_index and cufflinks are not displayed here because of limited space

2.2 Downloading and Installing Software

Please follow the corresponding instruction for each piece of software to install it or ask your computer administrator for help.

1. tRNAscan-SE: downloadable from http://lowelab.ucsc.edu/software/tRNAscan-SE.tar.gz and used for scanning tRNA genes from the genome sequence (*see* **Note 2** for alternative). The version number is 1.3.1.
2. FASTX-toolkit: downloadable from http://hannonlab.cshl.edu/fastx_toolkit/download.html (last accessed, 02/05/2017), used for preprocessing RNA-seq reads to eliminate low-quality ones.
3. Bowtie2 and TopHat: downloadable from http://bowtie-bio.sourceforge.net/bowtie2/index.shtml (last accessed, 02/05/2017) and https://ccb.jhu.edu/software/tophat/index.shtml (last accessed, 02/05/2017), and used for mapping RNA-seq reads to the genome.
4. Cufflinks: downloadable from http://cole-trapnell-lab.github.io/cufflinks/ (last accessed, 02/05/2017) and used for quantifying gene expression based on mapped reads.
5. CUA: available at CPAN http://search.cpan.org/dist/Bio-CUA/ (*see* **Note 9**) and can be installed using the following command in a command line:

```
$ cpan Bio::CUA
```

The above command assumes that the program PERL (https://www.perl.org/) has been installed. If not, please install PERL before installing CUA.

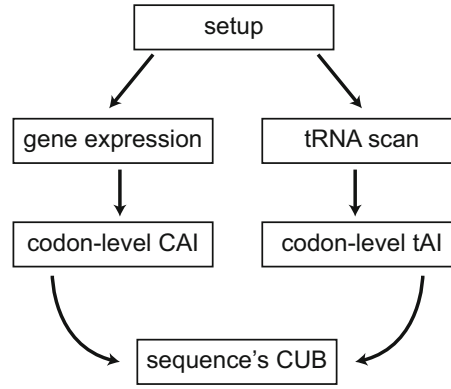

Fig. 2 The outline of running the protocol

3 Methods

Here we show the six main steps of calculating the four CUB metrics, CAI, tAI, Fop, and ENC (Fig. 2). Some steps may be skipped or modified, depending on whether the output has been obtained from other sources. For clarity, the commands typed in a command terminal start with the symbol "$", as shown above.

3.1 Setting Up the Environment

1. Create sub-folders under the folder "cub" according to the folder structure shown in Fig. 1, with the following commands:

```
$ cd cub
$ mkdir gene_exp
$ mkdir sequences
$ mkdir tRNA_scan
$ mkdir codon_cub
```

Then move, uncompress, and rename the downloaded files as follows:

```
$ gunzip -c Bdistachyon_283_v2.1.cds_primaryTrancsriptOnly.
fa.gz > sequences/cds.fa.
$ gunzip -c Bdistachyon_283_assembly_v2.0.softmasked.fa.gz
> sequences/genome.fa.
$ gunzip -c Bdistachyon_283_v2.1.gene_exons.gff3.gz > gen-
e_exp/exon.gff.
$ mv SRR921317.fastq.gz gene_exp.
```

3.2 Estimate Gene Expression

1. Preprocess RNA-seq reads to remove low-quality reads.
   (a) Convert the sequence quality scores in fastq files from Phred + 64 into Phred + 33 score using a custom Perl script "convert_fastq_quality.pl" (downloadable from https://github.com/fortune9/CUB_Brachypodium_distachyon). See this webpage https://en.wikipedia.org/wiki/FASTQ_format#Encoding for more information about the fastq format. Enter the folder "gene_exp", then run the following command

```
$ gunzip -c SRR921317.fastq.gz | convert_fastq_quality.pl –offset 31 -i - | gzip > SRR921317.fastq.gz.tmp
$ mv SRR921317.fastq.gz.tmp SRR921317.fastq.gz
```

   (b) Trim the low-quality bases and filter out the low-quality reads after trimming

```
$ fastq_quality_trimmer -Q33 -t 30 -z -l 20 -i SRR921317.fastq.gz -o SRR921317.QT.fastq.gz
$ fastq_quality_filter -Q33 -q 30 -p 80 -z -i SRR921317.QT.fastq.gz -o SRR921317.QT.QF.fastq.gz
```

   The above commands trim the nucleotides with quality score < 30 from the read ends and eliminate the reads shorter than 20 nucleotides after trimming. Reads that do not have a nucleotide quality score ≥30 in at least 80% of the positions are also removed. Please check the help documents of the programs for more details on the used options.

2. Map the Filtered Reads onto the Genome
   (a) Prepare the Bowtie2 index files

   Enter the folder "gene_exp," and run the following command.

```
$ mkdir bowtie_index
$ bowtie2-build ../sequences/genome.fa bowtie_index/genome
```

   This will create several Bowtie2 index files in the folder "bowtie_index", with the file extension ".bt2".
   (b) Map the RNA-seq reads onto the genome. In the following command, tophat2 uses 12 processors (specified by the option "-p") to map reads (*see* **Note 10**). One may modify it if necessary and check the help documentation for the usage of other options.

```
$ tophat2 -o tophat_out -p 12 -g 20 –library-type fr-unstranded -G exon.gff bowtie_index/genome SRR921317.QT.QF.fastq.gz &
```

This will result in a folder "tophat_out", which contains a bam file "accepted_hits.bam" storing the information of the mapped reads onto the genome.

2. Quantify gene expression using cufflinks. In the folder "cub," run the following commands:

```
$ cufflinks -o cufflinks_out -p 12 –library-type fr-unstranded -G exon.gff -b ../sequences/genome.fa -u -N tophat_out/accepted_hits.bam &
```

The output folder "cufflinks_out" contains multiple files. The file "genes.fpkm_tracking" has the information of gene expression level, and its columns 5 and 10 are cut out using the following command for subsequent analyses.

$ cut -f 5,10 cufflinks_out/genes.fpkm_tracking > cufflinks_out/genes.fpkm.

3.3 Scan tRNA Genes

1. Run tRNAscan-SE on the genome, enter the folder "tRNA_scan", run the following command:

```
$ tRNAscan-SE -o tRNA_scan.out ../sequences/genome.fa
```

which outputs the identified tRNA genes in the file "tRNA_scan.out".

2. Obtain the total number of tRNA genes (*see* **Notes 3** and **4**)for each anticodon with the following command:

```
$ cat tRNA_scan.out | sed -n -e '1,3!p' | tr -s ' ' '\t' \
$ | gawk '$5!="Pseudo"' | cut -f 6 | sort | uniq -c | tr -s ' ' '\t' \
$ | gawk 'BEGIN{OFS="\t"}{print $2, $1}' >anticodon_copy.tsv
```

The output is in "anticodon_copy.tsv" and pseudogenes are excluded in the calculation. The backslash "\" at the end of each line in the command is to connect the texts in multiple lines into one command. This is useful when the typed command is too long to fill into one line.

3.4 Codon-Level CAI

Enter the folder codon_cub, and run the following command to get the CAI value for each codon based on the top 300 highly expressed genes (*see* **Note 5**):

```
$ cai_codon.pl -i ../sequences/cds.fa –e ../gene_exp/cuf-flinks_out/genes.fpkm –s 300 –o codon_cai_top300.tsv
```

3.5 Codon-Level tAI

Enter the folder codon_cub, and run the following command (*see* **Note 6**).

```
$ tai_codon.pl -t anticodon_copy.tsv -o codon_tai.tsv
```

3.6 Get the Optimal Codons

Optimal codons may be defined in different ways. One popular way is to get codons with tAI higher than a threshold. However, the threshold is usually arbitrary, so the metric Fop based on determined optimal codons is less preferred than other metrics. Here, we define all codons with tAI greater than 0.47 as optimal codons. The following command can output the optimal codons:

```
$ gawk "$2 > 0.47" codon_tai.tsv > optimal_codons.tsv
```

3.7 Sequence-Level CUB

In this step, we calculate all the CUB metrics for each CDS sequence (*see* **Note 8**) using the following command after entering the folder "cub."

```
$ cub_seq.pl -s sequences/cds.fa -t codon_cub/codon_tai.tsv -c
codon_cub/codon_cai_top300.tsv -f codon_cub/optimal_codons.
tsv -enc -o seq_cub.tsv
```

The output file "seq_cub.tsv" contains several lines starting with "#" at the beginning, which provide information on how the data are arranged in the file. Figures 3 and 4 show the distribution of the four metrics calculated for the 30,626 genes and their correlations, respectively. As expected, *ENC* is negatively correlated with the other three metrics, and among the latter they are positively correlated.

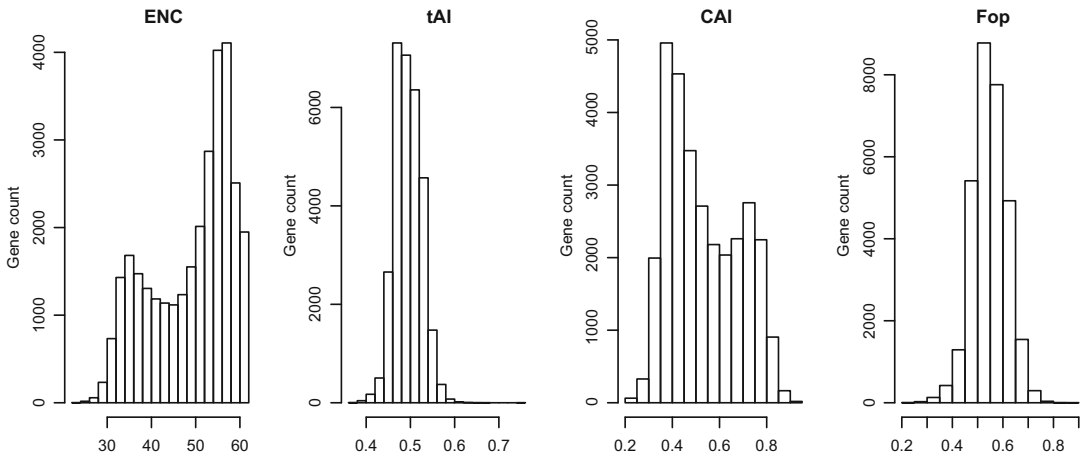

Fig. 3 The frequency distribution of the four CUB metrics

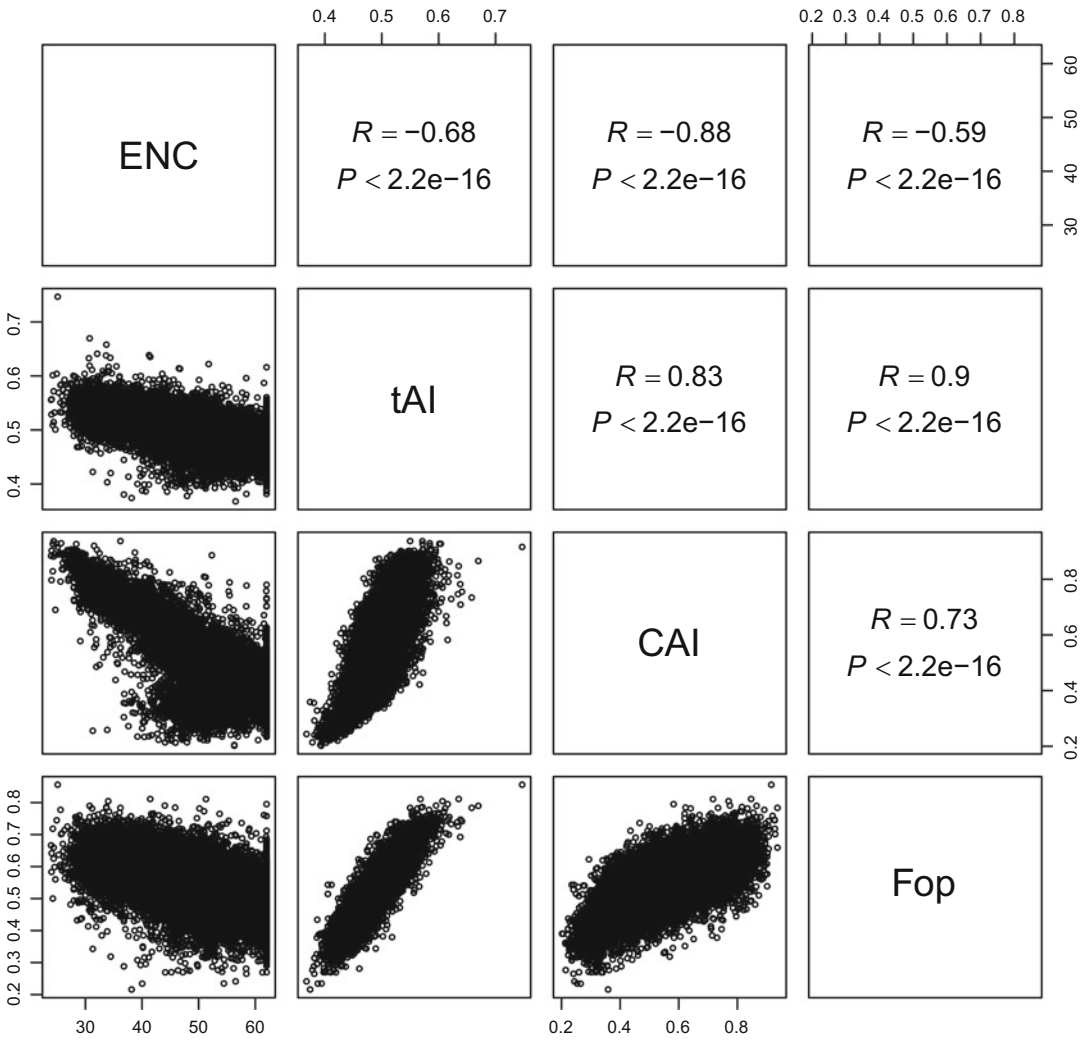

Fig. 4 The Pearson correlation among the four metrics

4 Notes

1. To download the data from the Phytozome website, one need register at the site first, which is free.
2. In this protocol, we use the tRNAscan-SE to indentify tRNA genes in the *Brachypodium distachyon* genome. For many species, the identified tRNA genes are already available in the genomic tRNA database http://gtrnadb.ucsc.edu/, so one may simply download the genes there and skip the relevant steps in the protocol.

3. At present, the tRNA gene copy number per anticodon is used to estimate the tRNA expression abundance. This estimate can be replaced with true tRNA abundance when available.
4. In the genomes of vertebrates, worms, and some plants, tRNA-derived repeats may be reported by tRNAscan-SE as tRNA genes. The genomic tRNA database suggests using the solution at http://gtrnadb.ucsc.edu/faq.html (last accessed on 02/05/2017) to filter out these repeat-derived tRNA genes.
5. Here the gene expression level measured in the 12-day-old seedlings is used to define highly expressed genes for estimating codons' CAI values. One can use the gene expression level from other biological samples for this purpose, but the computed metrics should be similar.
6. When calculating tAI, start codons and stop codons are generally excluded, and when calculating CAI, all codons for non-degenerate amino acids (Tryptophan W and Methionine M) and stop codons are excluded.
7. Beware the format of downloaded fastq files from NCBI. In most time, quality scores are encoded in the Phred + 33 format, but sometimes, the Phred + 64 format is used, just as the data in this paper. Since the Phred + 33 format is more acceptable by most programs, one need convert the formats when necessary.
8. Before feeding CDS sequences into any program, one may filter out the sequences without start codons, harboring internal stop codons, or incomplete as these problematic sequences may halt software.
9. Here we use CUA to calculate all CUB metrics. However, other programs may be used for certain circumstances. For example, codonW (http://codonw.sourceforge.net/) can be used for computing *ENC*, *CAI*, and *Fop* for a few species. codonR (http://people.cryst.bbk.ac.uk/~fdosr01/tAI/) can be used for calculating tAI in *Escherichia coli* whose tRNA gene copy numbers are enclosed in the program; for other species, one need specify tRNA gene copy number.
10. If analyzed files are large, one may think about running program in a parallel mode. For example, splitting a large file into small ones and run a program on each in parallel, or turn on multiple-cpu options such as –p in TopHat.

Acknowledgement

ZZ is grateful to Dr. Daven C Presgraves for the strong support in ZZ's research.

References

1. Vicario S, Moriyama EN, Powell JR (2007) Codon usage in twelve species of *Drosophila*. BMC Evol Biol 7:226
2. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci U S A 101(10):3480–3485
3. Bulmer M (1991) The selection-mutation-drift theory of synonymous codon usage. Genetics 129(3):897–907
4. Pop C, Rouskin S, Ingolia NT, Han L, Phizicky EM, Weissman JS, Koller D (2014) Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. Mol Syst Biol 10:770
5. Goodman DB, Church GM, Kosuri S (2013) Causes and effects of N-terminal codon bias in bacterial genes. Science 342(6157):475–479
6. Bentele K, Saffert P, Rauscher R, Ignatova Z, Bluthgen N (2013) Efficient translation initiation dictates codon usage at gene start. Mol Syst Biol 9:675
7. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J (2012) Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet 8(3):e1002603
8. Singh ND, Davis JC, Petrov DA (2005) X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis*. Genetics 171 (1):145–155
9. Hambuch TM, Parsch J (2005) Patterns of synonymous codon usage in *Drosophila* melanogaster genes with sex-biased expression. Genetics 170(4):1691–1700
10. Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1):32–42
11. Eyre-Walker AC (1991) An analysis of codon usage in mammals: selection or mutation bias? J Mol Evol 33(5):442–449
12. Akashi H (1994) Synonymous codon usage in *Drosophila* melanogaster: natural selection and translational accuracy. Genetics 136 (3):927–935
13. Zhang Z, Presgraves DC (2016) *Drosophila* X-linked genes have lower translation rates than autosomal genes. Mol Biol Evol 33 (2):413–428
14. Sharp PM, Li WH (1987) The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15 (3):1281–1295
15. dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32(17):5036–5044
16. Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146(1):1–21
17. Wright F (1990) The "effective number of codons" used in a gene. Gene 87(1):23–29

Chapter 12

Identification of Pseudogenes in *Brachypodium distachyon* Chromosomes

Salvatore Camiolo and Andrea Porceddu

Abstract

Pseudogenes are gene copies that have lost the capability to encode a functional protein. Based on their structure, pseudogenes are classified in two types. Processed pseudogenes arise by a process of retrotranscription from a spliced mRNA and subsequent integration into the genome. Nonprocessed (or duplicated) pseudogenes are generated by genomic duplication and subsequent mutations that disable their functionality so that they cannot longer encode a functional protein. Differently from processed pseudogenes, duplicated pseudogenes are expected to conserve the exon–intron structure of their functional paralogs.

Here, we describe a computational pipeline for identifying pseudogenes of both types in *B. distachyon* chromosomes. Our pipeline (1) identifies pseudogenes based on tBLASTn searches of *B. distachyon* proteins against the noncoding genomic complement of the same species, (2) identifies the most homologous pseudogenes functionally paralogous as the pseudogene paternal locus, (3) uses the intron–exon structure of paternal genes to distinguish between pseudogene types.

The pipeline is presented in its composing steps and tested on the *Brachypodium distachyon* Bd1 scaffold.

Key words Processed pseudogenes, Duplicated pseudogenes, Disabling mutations, Intron–exon structure, Paralogous genes

1 Introduction

Brachypodium distachyon is emerging as a model for functional and structural genomics of grasses [1]. It has a relatively small genome (~270 Mb) that results from a complex history of genome duplication, and chromosome fusions and reshuffling. Exhaustive analysis of transposable elements showed that retrotransposons cover 21.4% of the genome compared to 26% in rice, 54% in sorghum, and 80% in wheat [1]. Wicker et al. [2] suggested that transposable element dynamics and pseudogene evolution are important to explain some structural differences between wheat or barley and *B. distachyon* chromosomes. Other studies have highlighted the importance of retroposition to generate pseudogenes or chimeric genes in grasses [3]. Finally, an involvement of pseudogene derived

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_12,

transcripts in gene regulation has been demonstrated in rice by Guo et al. [4].

Although pseudogenes play a primary role in grass genome evolution, little information is to date available on pseudogene abundance and distribution in the grasses' model species *B. distachyon*.

Most of the pseudogenes identification protocols rely on two major steps: first, sequences featuring a genic architecture are identified within assumed not coding regions, and then their coding potential is investigated.

Unrevealing of "genic related" sequences is usually obtained by homology searches against known functional genes, aiming the identification of genes relics in intergenic or intronic regions [5]. Several pipelines for sequence homology searches have been developed [7]. For example, PseudoPipe [5] identifies pseudogenic regions based on tblastn searches that compare an amino-acidic query sequence against a nucleotide sequence database dynamically translated in all reading frames. The hits identified are then assembled in putative pseudogene loci based on the model of the original locus that generated them. The procedure is highly flexible as any sequence can be used for query searches; for example, Zhang et al. [6] have demonstrated that by querying mouse sequences that have no orthologous in *Homo sapiens* against *H. sapiens* genome it is possible to identify relics of genes (namely unitary pseudogenes) whose functional counterpart has been lost after the divergence of *H. sapiens*-rodents.

Other computational pipelines are based on searches of DNA query sequences against genomic sequences and use different strategies to assemble the hits in pseudogenic models (for a review *see* [7]).

In most cases the validation of the coding potential of candidate pseudogenes is obtained through the identification of disablements (i.e., stop codons or frameshifts) or through the analysis of the alignment with the cognate protein sequences.

Another approach for the pseudogenes identification is based on "ab initio" identification of genomic regions that, satisfying a series of conditions (i.e. presence of open reading frames, presence of splicing sites etc.), are predicted as genic models [8]. Unfortunately, most of these "ab initio" software programs have the tendency to deal with disablements by predicting introns or interrupting long ORFs. Direct consequences of this behavior are that while alternative pseudogene models are overlooked, genic models with unusual structure are predicted. In this regard Thibaud-Nissen et al. [8] have analyzed the coding potential of rice genic models with unusual features and found that a high proportion of these genes has either no associated transcript or show evidence of disablements. The histories of these pseudogenes were then recapitulated by homology searches with known rice

genes following the same procedures used, for this purpose, by homology-based pipelines.

Manual validation of genic predictions of *B. distachyon* genes belonging to cell wall biosynthesis and modifications and transcription factors belonging to sixteen families [1] have identified only few pseudogenes. Indeed the refinements of the *B distachyon* gene predictions by integrating, RNA seq data have highlighted the high accuracy of *B distachyon* genic predictions [1]. Less efforts have been dedicated to the identification of pseudogenes in regions predicted as noncoding.

In this chapter, we provide a detailed description of the bioinformatic procedure that can lead to the homology-based identification and classification of pseudogenes in *Brachypodium distachyon* chromosomes.

2 Materials

In this section, we list all files and software that are needed for pseudogene identification in *B. distachyon* intergenic and intron sequences.

2.1 Files

– ***B. distachyon*** **scaffold sequence file.**

This is a multifasta file that can be downloaded from http://www.ncbi.nlm.nih.gov/genome/?term=Brachypodium+distachyon (last accessed on 30th January 2017) or from the phytozome website (http://phytozome.jgi.doe.gov/pz/portal.html, last accessed on 30th January 2017).

In this chapter, we refer to the hard masked scaffolds that can be downloaded from the Phytozome website and report repetitive sequences masked.

– ***B. distachyon*** **gff file**.

This is a gff formatted annotation file (http://www.ensembl.org/info/website/upload/gff.html, last accessed on 30th January 2017) listing all genic predictions relative to the scaffold sequences. We used the file Bdistachyon_314_v3.1.gene.gff3 obtained from the Phytozome website.

– ***B. distachyon*** **proteins multifasta file**.

A query file listing the *B. distachyon* protein sequences was used to perform the procedure RET (Fig. 1). Such a file can be downloaded from the phytozome ftp site or independently generated with the aid of a dedicated software. We used gff2sequence (v 0.2) [10] software but also other software produce suitable results.

A query file listing the exon derived peptides was used in order to perform the procedure DUP. This file is generated in a two-step process. First exons were tailed on both ends by adding a number of adjacent nucleotides (tailed exons). In this chapter we used tails of

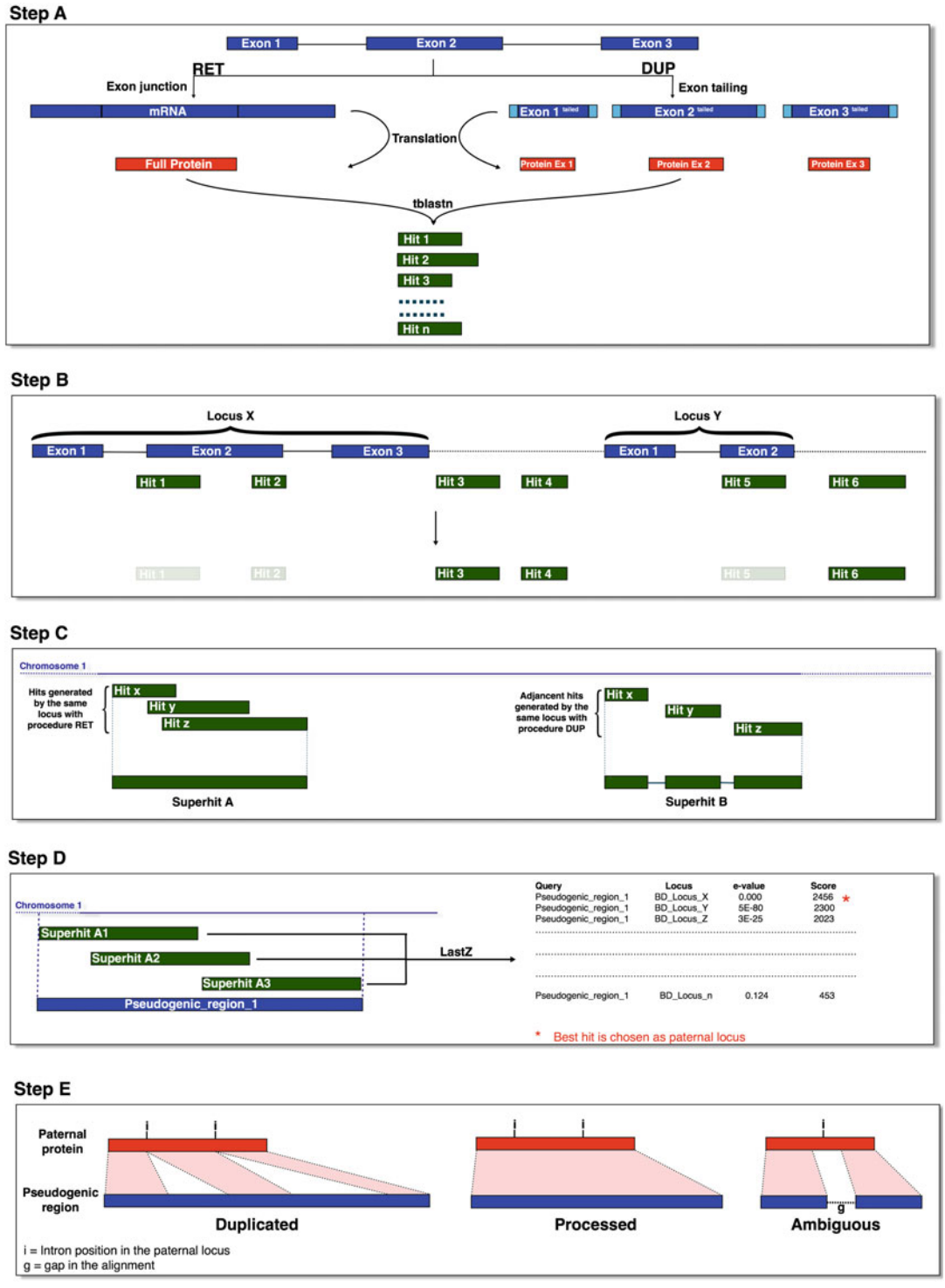

Fig. 1 Outline of the pipeline used to identify pseudogenes

51 nucleotides (a perl script for generating this files is available upon request to the authors). Additional nucleotides were added in according to the exon ending/starting with the first, second or third codon nucleotide in order to avoid truncated codons in the tailed exons (i.e., an additional two nucleotides were added to the end of the tailed exon if its last nucleotide represented the first base of the codon). Tailed exons were then converted into amino-acid sequence.

– ***B. distachyon*** **cds file**.

A multifasta file listing all *B. distachyon* cds sequences. Such file can be downloaded from phytozome web site.

– ***B. distachyon*** **genomic loci sequences**.

A multifasta file listing all genomic sequences corresponding to primary transcripts. This file can be generated by gff2sequence [10].

– ***B. distachyon*** **Gene Ontology**.

A file reporting the gene ontology for *B. distachyon* loci. We used the Bdistachyon_314_v3.1.annotation_info.txt file from the Phytozome 11 website.

This a tab-delimited file reporting the following fields.

1: Phytozome internal transcript ID, 2: Phytozome gene locus name, 3: Phytozome transcript name, 4: Phytozome protein name, 5: PFAM, 6: Panther, 7: KOG, 8: KEGG ec, 9: KEGG Orthology, 10: Gene Ontology terms, 11: best Athaliana TAIR10 hit name, 12: best Athaliana TAIR10 hit symbol, 13: best Athaliana TAIR10 hit defline, 14: best Osativa v7.0 hit name. 15: best Osativa v7.0 hit symbol, 16: best Osativa v7.0 hit defline.

2.2 Software

Users should have blastall (v.2.2.26) [9], fasta36 (v. 36.3.8c) [11], LASTZ (v. 1.02.00) [12] and Genewise (v. 2.4.1) [13] executables in their PATH. All these software are freely available and can be installed following the instructions reported in the corresponding documentation.

3 Methods

3.1 Pipeline Overview

The pipeline for pseudogene identification (Fig. 1) was inspired by PseudoPipe [5]. However, since each step was independently developed a perfect mirroring of the produced results is not to be expected. Hereafter we report the steps involved in the pipeline.

Step A. Putative pseudogenes were identified by tblastn [9] searches with the *B. distachyon* scaffold sequences as a subject and either (1) exon derived peptide sequences (Procedure DUP in Fig. 1) or (2) the full length protein sequence (Procedure RET in Fig. 1) as queries.

Step B. Only the hits not overlapping predicted genic regions were considered for further analyses.

Step C. Hits filtered as explained above and that were identified by the same query sequence were merged into "superhits" (Procedure RET in Fig. 1). Accordingly, hits that were generated by queries corresponding to sequential exons of the same genic model were merged into superhits if the intervening distance was smaller than a given threshold (Procedure DUP in Fig. 1). Such a threshold was dynamically determined considering the length of the intron between the two exons in the query gene model.

Step D. Hits that overlapped for more than 20% of their length were grouped together. For each of these groups we reanalyzed the alignments between the pseudogenes and the query sequences at DNA level. The best matching pair pseudogene-query sequences for each group were chosen. This selection was based on the hypothesis that under neutral evolution the pseudogene remains more similar to the modern form of the locus that generated it than to the sequence of the paralogs.

Step E. Pseudogenes were classified as duplicated (the parental intron–exon structure is maintained), processed (no similarity to intronic sequences is observed due to messenger RNA mediated retroposition) or ambiguous (the absence of sequence information prevents a precise classification) by studying protein alignment to the query sequence.

3.2 Pipeline Protocols

3.2.1 tblastn Searches

Step A is performed by first indexing the hard masked *B. distachyon* genome with the following command:

```
makeblastdb -in B.distachyon_283_assembly_v2.0-hardmasked.fa
-dbtype nucl.
```

followed by the tblastn search (*see* Pipeline overview):

```
tblastn -query proteins.fasta -db
Bdistachyon_283_assembly_v2.0-hardmasked.fa -outfmt 6 -out
Results.txt.
```

proteins.fasta is the name of the query file that was used in this work and contains either the full length proteins (procedure RET) or the translated tailed exons (procedure DUP) in a multifasta format as explained in detail in the Pipeline overview section. The results are stored in the Results.txt file in a tabular format (-outfmt 6). Additional flags can be added to the tblastn command in accordance to the blast [9] user manual. As a way of example, a multiprocessor elaboration can be performed by adding the -num_threads flag). Moreover, to limit the size of Results.txt file it is possible to filter the matches with low similarity by setting a threshold in terms of e-value or percentage of identity (flags -evalue and -perc_identity, respectively). However, since filtering at this stage is not reversible, and taking into account that tblastN search represents the most time-consuming step, we discourage the use of

such constraints at this stage but rather filter the obtained hits in a following step (see below).

3.2.2 Filtering the Results

Several filters may be used at this step to remove the hits that do not meet some user-defined criteria. We adopted filters on percentage of identity ($>= 40\%$) and probability of obtaining them by chance (e-value below 0.0001) while classifying the identified hits based on the strand and genomic scaffold. Note that at this stage, the filtered hits include also alignments between query sequences and annotated transcripts. However, only those hits that do not overlap annotated cds regions should be kept for further analysis. We usually perform this task by checking that the hit maps (without overlap) in noncoding regions. Attention may be required in performing this action with gff files that include overlapping features that should be merged beforehand. An index was attributed to each filtered hit to allow information retrieval at later steps.

3.3 Hits Merging

At this step, overlapping or adjacent hits are merged following user-defined criteria. First, the hits are grouped based on the matching query sequence; then overlapping hits within each group are merged to form super-hits (*see* Case I in Table 1).

Adjacent (super)hits that matched the same query sequence were further merged if their distance on the chromosome was low compared to the corresponding distance of matching sequences in the query. In practice, these hits were considered to belong to the same pseudogene and the intervening gap generated by (1) low complexity or very decayed regions, (2) short insertions within pseudogene, and (3) short repetitive elements. The criterion we adopted is: two hits that are separated by a distance Gs and identified by adjacent regions of a given query at distance Gq are considered eligible to merging, thus generating a superhit if Gs $<100 + 3 \times$ Gq. An example of such a case is reported in Table 1 (Case II).

An additional and more complex situation may arise when "tailed translated exons" are used as queries (Procedure DUP). Adjacent exons of the same locus may identify adjacent hit in the scaffold sequence (case III in Table 1).

Finally, independent events may be inferred for hits that are found at distance exceeding a given threshold. We merged super-(hits) identified by adjacent query (exon derived) sequences if the size of the putative pseudo_intron (i.e., the distance between hits) and that of the corresponding intron in the matching query locus differed for not more than 500 bp. A difference between putative pseudo-intron and the corresponding intron from the functional locus is needed to take into account possible insertion or deletion occurring in the pseudogene.

Because in most cases, the applied (similarity) thresholds for hit filtering does not identify unique query to pseudogenic sequence

Table 1
Hits merging

Query ID	Subject ID	Perc. identity	Aln length	Mismatch count	Gap open count	query start	query end	Subject start	Subject end	Eval	Score	Hit index
Case I												
Bradi1g68800.1._1	Bd1	71.62	74	19	2	172	243	67,569,260	67,569,481	8.00*E*−24	98.2	113,246
Bradi1g68800.1._1	Bd1	65	20	7	0	261	280	67,569,496	67,569,555	8.00*E*−24	27.7	113,247
Bradi1g68800.1._1	Bd1	66.67	15	5	0	274	288	67,569,537	67,569,581	8.00*E*−24	25	113,248
Case II												
Bradi1g07195.1_1	Bd1	65.38	26	9	0	23	48	42,202,851	42,202,928	3.00*E*−06	40.4	10,419
Bradi1g07195.1_1	Bd1	77.42	31	7	0	50	80	42,202,936	42,203,028	3.00*E*−06	28.1	10,420
Case III												
Bradi1g40701.1_2	Bd1	94.23	52	3	0	13	64	42084475	42084630	2.00*E*−24	100	70268
Bradi1g40701.1_2	Bd1	61.11	36	14	0	14	49	53966204	53966311	8.00*E*−07	48.5	70270
Bradi1g40701.1_3	Bd1	93.85	65	4	0	12	76	42084929	42085123	7.00*E*−35	125	70273
Bradi1g40701.1_3	Bd1	100	14	0	0	1	14	42084897	42084938	7.00*E*−35	40.4	70274
Bradi1g40701.1_4	Bd1	94.44	54	3	0	75	128	42085855	42086016	2.00*E*−40	108	70279
Bradi1g40701.1_4	Bd1	95	40	2	0	35	74	42085736	42085855	2.00*E*−40	76.3	70280
Bradi1g40701.1_4	Bd1	50	70	35	0	35	104	53967300	53967509	1.00*E*−14	72.4	70281

Bradi1g40701.1_4	Bd1	55.56	18	8	0	18	35	53967250	53967303	1.00*E*−14	25	70282
Bradi1g40701.1_4	Bd1	68.89	45	14	0	3	47	42085641	42085775	1.00*E*−12	68.2	70283
Bradi1g40701.1_4	Bd1	50	42	20	1	40	81	12579493	12579615	5.00*E*−06	42	70284
Bradi1g40701.1_4	Bd1	63.16	19	7	0	17	35	12579425	12579481	5.00*E*−06	25.8	70285
Bradi1g40701.1_5	Bd1	94.12	51	3	0	1	51	42086004	42086156	1.00*E*−22	94.4	70287
Bradi1g40701.1_6	Bd1	78.79	33	7	0	1	33	42086131	42086229	1.00*E*−18	56.6	70289

Three cases of hits merging are presented. I and II refer to the merging of hits matching the same query sequences. Different hits matching adjacent exons of the same transcript are merged in III

Case I: Hits 113,246, 113,248, 113,247 are merged in a SuperHit because overlapping. The resulting hit is: Bradi1g68800.1_1 in range 67,569,260–67,569,581, identified by hits 113,246–113,247–113,248

Case II: The hits 10,419 and 10,420 matching the same query Bradi1g07195.1_1 are merged though not overlapping. The distance between hits (Gs) is 8 bp and the distance (Gq) in the matching query sequence is 6 [=(50–48) × 3]. The resulting Super(Hit) is: Bradi1g07195.1_1 in range 42,202,851–42,203,028 identified by hits 10,419–10420

Case III: the hits 70628,70274, 70283–70280, 70287, 70289–70290 are merged in a superhit because identified by adjacent introns of locus Bradi1g40701.1. The resulting SuperHits are: Bradi1g40701.1 _2 in range 42084475–42084630 identified by hit 70268; Bradi1g40701.1_3 in range 42084897–42084938 identified by hit 70274; Bradi1g40701.1_4 in range 42085641–42085855 identified by hits 70283–70280; Bradi1g40701.1_5 in range 42086004–42086156 identified by hit 70287; Bradi1g40701.1_6 in range 42086131 42086303 identified by hits 70289–70290. The other hits are positioned at a distance that is not compatible with the corresponding intron length and will give rise to other superhits

Table 2
Step D: Pseudoregion identification

	Query ID	Subject ID	Subject start	Subject end	Hit index
Pseudoregion 1	Bradi2g34986.1_2	Bd1	77213	77578	160673
	Bradi4g09312.1_2	Bd1	77213	77581	266253
	Bradi4g09530.1_2	Bd1	77222	77575	277422
	Bradi4g09648.1_1	Bd1	77222	77581	277663
	Bradi4g09271.1_1	Bd1	77222	77581	266230
	Bradi4g09631.1_2	Bd1	77222	77581	277644
	Bradi4g10045.1_2	Bd1	77231	77566	278152
	Bradi4g10045.2_1	Bd1	77231	77566	278150
	Bradi4g10040.1_2	Bd1	77243	77566	278147
	Bradi4g10197.1_1	Bd1	77249	77491	278183
	Bradi4g10220.2_1	Bd1	77255	77566	278194
	Bradi2g03200.2_1	Bd1	77270	77455	129033
	Bradi2g03260.1_1	Bd1	77270	77566	129040
	Bradi1g00278.1_4	Bd1	77393	77581	1188
Pseudoregion 2	Bradi4g09648.1_1	Bd1	89749	90858	277658
	Bradi4g09631.1_2	Bd1	89749	90858	277641
	Bradi2g34986.1_2	Bd1	89749	90870	160671
	Bradi4g09271.1_1	Bd1	89749	90873	266225
	Bradi4g09530.1_2	Bd1	89749	90873	277417
	Bradi4g10220.2_1	Bd1	89761	89976	278196
Pseudoregion 3	Bradi1g00278.1_5	Bd1	90196	91009	1195–1196
	Bradi4g09511.1_2	Bd1	90616	90870	277391
	Bradi2g52840.1_1	Bd1	90866	90916	183983
	Bradi2g51807.1_1	Bd1	90866	90922	181337
	Bradi3g14917.1_2	Bd1	90866	90928	213740
	Bradi3g14917.2_1	Bd1	90866	90928	213738
	Bradi4g11940.2_3	Bd1	90866	90991	281090

(Super)-Hits identified as in Step B-C are assigned to pseudogenic region based on overlapping coordinates. A (Super)-Hit is assigned to a pseudogenic region if its length overlaps for at least one fifth the last added hit of the region.

3.4 Determining Paternity of Pseudogenes

relationship, the ambiguities in paternal locus identification must be resolved. Overlapping (super)hits from previous steps are grouped together in a list that identify a so called pseudogenic region. Such lists include all query-(super)hit relations that possibly relate to a single pseudogenization event. A region of overlap between two sequences is considered significant if it covers at least one fifth of the length of both sequences. Note that here we consider transitivity in overlapping, i.e., if A1 and A2 overlap and A1 overlaps A3 than A1, A2, and A3 all overlaps and are attributed to the same list (an example of pseudogenic regions is reported in Table 2).

For each pseudogenic region we studied the alignment of the corresponding genomic sequences (the putative pseudogene) to

the original query locus. The putative paternal locus, i.e., the locus that generated the pseudogene, is inferred based on the score of its alignment to the pseudogene sequence. We suggest to use LASTZ for alignments of pseudogenes to the genomic sequences of query locus. In alternative it is possible to study the outputs of tblastN of all the hits matching a given pseudogenic regions. The paternal locus in this case is identified by comparing the e-value (before merging) and score of hits identified by tblastn. The latter approach is faster but does not take into account the alignment of intron (from the functional locus) to putative pseudo-intron sequences.

3.5 Pseudogene Classification

Pseudogenes are thought to be generated by duplication and subsequent disablement of coding potential, or by transposition (retroposition) of an mRNA retro-transcribed sequence. Hence, several genomic features can be used to discriminate between these two types. The presence of sequence showing high similarity to intron sequence of the "paternal locus" are diagnostic of a duplication based event and the presence of disablement such as stop codons or frameshift mutation accounts for the inability to encode a functional protein. Conversely, the absence of sequences showing similarity to intron sequences, and the presence of a poly-adenine tail at the 3′ end are the main features used to predict a processed pseudogene.

Some of these diagnostic features in the putative pseudogene can be deduced from alignments to the query protein carried out by tfasty or Genewise. tfasty compares a query protein sequence to the pseudogene nucleotide sequence calculating similarities with frameshifts to the reverse and forward orientations while allowing frameshifts between and within codons [14].

Genewise was originally developed to refine gene predictions and allow the detection of intron exon splice sites in alignment of protein sequences to genomic loci. As this algorithm takes into account disablement it can be used also for pseudogene alignments.

Figure 2 reports a pseudogene analysis carried out on the output of Genewise. The genomic region 31,178,398–31,179,685 was identified as having high similarity to four adjacent exons of locus Bradi4g11118.1. The identity of matching query sequence are listed in section Q while the exon intron structure of the paternal locus is listed in section LS (Locus Structure). The Bradi4g11118.1 has four introns: the first at position 2 (phase 0), another at position 112 (phase 2), the third at position 148 (phase 0), and the last at position 228 (phase 1). Only intron at position 112 is predicted by Genewise (note row GW_predicted in C section of Fig. 1). The classification of the other intron positions is obtained considering the data reported in Alignment Parsing (AP) section. The Bradi4g11118.1 protein is aligned to the

Fig. 2 Pseudogene-functional protein alignment. An example of analysis of the alignment obtained with Genewise along with the classification of exon–intron junction of the pseudogene and other features

putative pseudogene derived peptide from aa 5 to 228 (see Cover in AP). Twelve consecutive gaps are found starting from position 148 (as annotated in Coverpf). Finally, gaps are present in the pseudoprotein starting from position 1 to 5 (these positions are referred to the Bradi4g11118.1 protein sequence). Exon–intron junction classification is carried out in four phases. At phase 1, it is checked whether there are correspondences between intron positions

detected by Genewise and locus structure. Whether a region with gaps is found in correspondence of a position where introns may be expected based on locus structure is checked at phase 2. An example in case is offered by the second exon–intron junction (position 148). Twelve gaps are found at position 148 where an intron is predicted for the Bradi4g11118.1 locus. Indeed this position has been classified as DUP_amb to remark that it was not recognized as a true intron by Genewise. The exon–intron at position 2 is classified as ambiguous (Amb) because, although the query of the first exon was included in the superhit, there is no pseudogene sequence information in the alignment suitable for classification (note that the presence of gaps are reported in the Gap_p of the AP section). Finally the alignment does not cover adequately the region where the fourth intron is expected and then no classification for a such intron is attempted. The output reports also the gene ontology (GO) of the functional locus and the presence of disablements and poly adenine tails.

For polyadenine detection we surveyed a region of 600 bp that was 3′ to the pseudogene region, with a sliding window of 50 nucleotides. The size of the surveyed region (600 bp) was chosen because more that 85% of *B. distachyon* 3′-utr are shorter that this size. A poly Adenine tract was identified as a sequence window with more than 30 adenines.

4 Notes

We used the DUP pipeline to identify pseudogenes in the *B. distachyon* chromosome (Bd1 scaffold). 13,985 pseudogenic regions were identified (5770 with plus and 8215 with minus orientation). Ambiguities in paternal locus identities were resolved based on homology score between pseudogene and genomic sequence using LASTZ, or by comparing the *e*-value (before merging) and score of hits identified by tblastn. Table 3 lists the locations of 60 *B. distachyon* pseudogenes located in Bd1 scaffold along with information about their classification and presence of disablements or poly Adenine sequence at the 3′ pseudogene sequence.

Table 3
Examples of pseudogenes identified on *B. distachyon* chromosome 1. Dup stands for duplicated pseudogenes. Amb for ambiguous pseudogene

Pseudogene coordinates					Pseudogene features			Paternal locus		
Strand	Start	End	Type	Exons	Disablement	Poly-A	ID	Number exon	Locus_before	Locus_after
–	5E + 07	46249658	Dup	3	Yes	NO	Bradi3g28811.1		Bradi1g47470.1	Bradi1g47480.1
–	1E + 06	961672	Dup	2	NO	NO	Bradi1g01450.1		Bradi1g01470.1	Bradi1g01470.1
–	1E + 07	12408102	Amb							
–	2E + 07	18415772	Dup	1	NO	NO	Bradi4g13222.1		Bradi1g22900.1	Bradi1g22907.1
–	4E + 07	40505125	Dup	5	YES	NO	Bradi3g10850.1		Bradi1g42873.1	Bradi1g42880.1
–	3E + 07	25241751	Dup	3	NO	NO	Bradi3g14710.1		Bradi1g29669.1	Bradi1g29690.1
–	1E + 05	127418	Amb	1						
–	6E + 07	55322629	Dup	3	Yes	NO	Bradi1g47460.1	3	Bradi1g56416.1	Bradi1g56458.1
–	6E + 06	6255252	Dup	4	No	NO	Bradi1g09140.1	7	Bradi1g08870,1	Bradi1g08875.1
–	5E + 07	47991229	Dup	3	Yes	No	Bradi1g47460.1	3	Bradi1g49010.1	Bradi1g49015.1
–	5E + 07	50255514	Dup	2	No	No	Bradi3g16827.1	2	Bradi1g51595.1	Bradi1g51598.1
–	5E + 07	50739214	Dup	2	Yes	No	Bradi4g02308.1	3	Bradi1g52120.1	Bradi1g52140.2
–	3E + 07	31192043	Amb	2	Yes	No	Bradi3g28811.1	4	Bradi1g35496.1	Bradi1g35500.1
–	5E + 07	49060240	Amb	2	No	No	Bradi4g02308.1	3	Bradi1g50147.1	Bradi1g50147.2
–	4E + 07	42863160	Amb	1	No	No	Bradi3g06810.1	3	Bradi1g44440.1	Bradi1g44460.1
–	2E + 05	152100	Amb	1	No	No	Bradi4g09271.1	4	Bradi1g00272.1	Bradi1g00278.1
–	3E + 07	32900976	Amb	1	No	No	Bradi1g10688.1	8	Bradi1g36971.1	Bradi1g36976.1
–	5E + 07	45624397	Dup	3	No	No	Bradi2g55860.1	4	Bradi1g46850.1	Bradi1g46860.1
–	6E + 07	58677883	Dup	4	Yes	No	Bradi1g69520.1	20	Bradi1g59300.1	Bradi1g59310.2
–	4E + 06	3821808	Dup	3	Yes	No	Bradi5g18560.1	6	Bradi1g05640.2	Bradi1g05650.1

–	$5E+07$	46168776	Dup	3	No	No	Bradi1g43151.1	4	Bradi1g47407.2	Bradi1g47427.2
–	$3E+07$	25277203	Amb	1	No	No	Bradi1g29693.1	1	Bradi1g29693.1	Bradi1g29697.1
–	$6E+07$	62059207	Dup	2	No	No	Bradi1g41291.1	2	Bradi1g62547.1	Bradi1g62576.1
–	$2E+07$	17062974	Amb	2	Yes	No	Bradi2g10656.1	3	Bradi1g21177.1	Bradi1g21190.1
–	$6E+07$	55066910	Dup	2	No	No	Bradi1g28323.1	2	Bradi1g56095.1	Bradi1g56100.3
–	$2E+07$	19805409	Dup	3	Yes	No	Bradi2g57850.1	3	Bradi1g24516.1	Bradi1g24528.1
–	$4E+07$	35181428	Amb	1	Yes	No	Bradi1g58105.1	2	Bradi1g38730.1	Bradi1g38735.1
–	$3E+07$	34024311	Amb	1	No	No	Bradi1g58105.1	2	Bradi1g37870.1	Bradi1g37876.1
–	$6E+07$	64389453	Dup	2	Yes	No	Bradi1g21440.1	9	Bradi1g64957.1	Bradi1g64970.1
–	$6E+07$	62891361	Amb	2	No	No	Bradi4g27965.1	2	Bradi1g63470.1	Bradi1g63480.1
–	$3E+07$	34766848	Amb	2	No	No	Bradi2g44210.1	3	Bradi1g38350.1	Bradi1g38353.1
–	$2E+07$	15657759	Amb	1	No	No	Bradi1g19580.1	4	Bradi1g19600.1	Bradi1g19607.1
–	$4E+07$	41328507	Dup	3	Yes	No	Bradi3g22684.1	7	Bradi1g43380.1	Bradi1g43990.1
–	$4E+06$	3823189	Dup	2	No	No	Bradi1g01670.1	11	Bradi1g05640.2	Bradi1g05650.1
–	$2E+07$	24554343	Amb	2	No	No	Bradi4g06141.1	4	Bradi1g29025.1	Bradi1g29030.1
–	$1E+06$	1324621	Amb	1	Yes	No	Bradi1g01950.1	1	Bradi1g01957.1	Bradi1g01965.1
–	$7E+07$	72330813	Amb	1	Yes	No	Bradi1g75065.2	1	Bradi1g75060.1	Bradi1g75065.1
–	$3E+07$	26069595	Amb	1	No	No	Bradi3g16827.1	2	Bradi1g30678.1	Bradi1g30690.1
–	$3E+07$	26758379	Dup	2	Yes	No	Bradi4g22988.1	6	Bradi1g31193.4	Bradi1g31200.1
–	$5E+07$	54223947	Amb	1	Yes	No	Bradi2g19723.1	1	Bradi1g55490.1	Bradi1g55500.1
–	$3E+07$	27522146	Amb	1	No	No	Bradi2g23660.1	1	Bradi1g32050.1	Bradi1g32060.1
–	$6E+07$	57625762	Amb	1	Yes	No	Bradi1g28135.1	1	Bradi1g58545.1	Bradi1g58550.1
–	$1E+07$	12729659	Amb	1	Yes	No	Bradi1g24126.1	1	Bradi1g15870.1	Bradi1g15790.1

(continued)

Table 3
(continued)

Pseudogene coordinates					Pseudogene features			Paternal locus		
Strand	Start	End	Type	Exons	Disablement	Poly-A	ID	Number exon	Locus_before	Locus_after
–	6E + 07	61635895	Dup	2	No	No	Bradi2g36506.1	2	Bradi1g62057.2	Bradi1g62063.1
–	5E + 07	47699807	Dup	3	No	No	Bradi2g04706.1	3	Bradi1g48813.1	Bradi1g48817.1
–	2E + 07	22505669	Amb	1	No	No	Bradi4g13916.1	2	Bradi1g27388.1	Bradi1g27400.2
–	2E + 07	23441048	Dup	2	No	No	Bradi1g41720.1	2	Bradi1g28230.2	Bradi1g28260.2
–	4E + 06	4345528	Amb	1	No	No	Bradi1g28185.1	2	Bradi1g06470.1	Bradi1g06470.2
–	2E + 07	24100257	Amb	1	No	No	Bradi2g08380.1	1	Bradi1g12697.1	Bradi1g12701.1
+	3E + 07	26160537	Dup	6	Yes	Yes	Bradi5g09708.1	6	Bradi1g30750.1	Bradi1g30767.1
+	2E + 07	23559726	Dup	8	Yes	Yes	Bradi3g24672.1	8	Bradi1g28323.1	Bradi1g28326.1
+	3E + 07	29522422	Dup	6	Yes	Yes	Bradi4g22988.1	6	Bradi1g33850.1	Bradi1g33860.1
+	4E + 07	36288639	Dup	6	Yes	No	Bradi4g17405.1	6	Bradi1g39520.1	Bradi1g39620.1
+	2E + 07	24905024	Dup	4	Yes	No	Bradi1g43455.1	4	Bradi1g29340.1	Bradi1g29350.1
+	4E + 07	40474711	Dup	2	Yes	Yes	Bradi1g39454.1	3	Bradi1g42850.1	Bradi1g42860.1
+	2E + 07	23186333	Dup	9	Yes	Yes	Bradi1g28000.1	9	Bradi1g28000.1	Bradi1g28010.1
+	4E + 07	36289111	Dup	8	Yes	Yes	Bradi3g24672.1	8	Bradi1g39520.2	Bradi1g39620.1
+	1E + 07	9622087	Dup	4	Yes	Yes	Bradi1g14040.1	6	Bradi1g12470.2	Bradi1g12750.1
+	3E + 07	26912127	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g31326.1	Bradi1g31330.1
+	2E + 07	22916689	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g27780.1	Bradi1g27790.1
+	6E + 07	57239061	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g58105.1	Bradi1g58110.1
+	6E + 07	56290963	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g57271.1	Bradi1g57280.1

+	$7E+07$	65140006	Dup	1	Yes	No	Bradi1g65990.1	1	Bradi1g65961.1	Bradi1g65964.1
+	$8E+06$	7855339	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g10795.1	Bradi1g10800.1
+	$3E+07$	32223629	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g36430.1	Bradi1g36441.1
+	$8E+06$	8422257	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g11357.1	Bradi1g11365.1
+	$3E+07$	28532797	Dup	4	Yes	Yes	Bradi1g05202.1	4	Bradi1g32890.1	Bradi1g32901.1
+	$2E+07$	17394937	Dup	3	Yes	No	Bradi2g37365.1	3	Bradi1g21560.4	Bradi1g21570,1
+	$4E+07$	43245173	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g44900.1	Bradi1g44920.1
+	$4E+07$	35653766	Dup	3	Yes	Yes	Bradi3g23840.1	3	Bradi1g39056.1	Bradi1g39056.1
+	$3E+07$	29443136	Dup	2	Yes	Yes	Bradi3g39278.1	3	Bradi1g33800.4	Bradi1g33803.1
+	$3E+07$	32298801	Dup	3	Yes	Yes	Bradi1g21002.1	3	Bradi1g36505.1	Bradi1g36520.1
+	$1E+07$	10871180	Dup	3	Yes	Yes	Bradi3g23840.1	3	Bradi1g13990.1	Bradi1g14000.1
+	$2E+07$	18717602	Dup	9	Yes	Yes	Bradi1g21130.2	10	Bradi1g23293.1	Bradi1g23299.1
+	$1E+07$	13061103	Dup	3	Yes	Yes	Bradi3g20965.1	4	Bradi1g16120.1	Bradi1g16140.1
+	$6E+06$	5702344	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g08080.1	Bradi1g08094.2
+	$8E+06$	8500566	Dup	2	Yes	Yes	Bradi1g39454.1	3	Bradi1g11430.1	Bradi1g11435.1
+	$5E+07$	50030354	Dup	2	Yes	Yes	Bradi1g39454.1	3	Bradi1g51345.1	Bradi1g51350.1
+	$2E+07$	24947915	Dup	3	Yes	Yes	Bradi3g20965.1	4	Bradi1g29356.1	Bradi1g29360.2
+	$3E+07$	34732132	Dup	3	Yes	Yes	Bradi4g22988.1	6	Bradi1g38330.1	Bradi1g38340.1
+	$4E+06$	4269990	Dup	1	Yes	Yes	Bradi4g20543.1	1	Bradi1g06372.1	Bradi1g06380.2
+	$3E+07$	27121009	Dup	1	Yes	Yes	Bradi4g20543.1	1	Bradi1g31538.1	Bradi1g31547.1
+	$5E+07$	54199818	Dup	3	Yes	Yes	Bradi3g23840.1	3	Bradi1g55480.1	Bradi1g55485.1
+	$4E+07$	36262419	Dup	3	Yes	Yes	Bradi1g05202.1	3	Bradi1g39490.1	Bradi1g39510.1
+	$3E+07$	27836751	Amb	2	Yes	Yes	Bradi3g18720.2	3	Bradi1g32430.1	Bradi1g32437.1

(continued)

Table 3
(continued)

Pseudogene coordinates					Pseudogene features			Paternal locus		
Strand	Start	End	Type	Exons	Disablement	Poly-A	ID	Number exon	Locus_before	Locus_after
+	4E + 07	36561406	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g39720.1	Bradi1g39920.1
+	1E + 07	14050721	Dup	2	Yes	Yes	Bradi5g10831.1	2	Bradi1g17520.1	Bradi1g17526.1
+	1E + 05	103382	Amb	1	Yes	Yes	Bradi4g09312.1	6	Bradi1g00237.1	Bradi1g00247.2
+	3E + 07	27928278	Amb	1	Yes	Yes	Bradi3g23840.1	3	Bradi1g32475.1	Bradi1g32480.1
+	6E + 07	59428564	Dup	2	Yes	Yes	Bradi3g28811.1	4	Bradi1g60010.1	Bradi1g60021.1
+	1E + 06	1182312	Amb	1	Yes	Yes	Bradi3g23840.1	3	Bradi1g01790.3	Bradi1g01795.1
+	4E + 07	41143466	Amb	1	Yes	Yes	Bradi1g43300.1	2	Bradi1g43271.1	Bradi1g43280.1
+	4E + 07	43020593	Dup	3	Yes	Yes	Bradi1g45085.1	5	Bradi1g44640.2	Bradi1g44640.4
+	89749	90873	Amb	1	Yes	Yes	Bradi4g09271.1	4	Bradi1g00232.1	Bradi1g00237.2
–	55321540	55322629	Dup	3	Yes	No	Bradi1g47460.1	3	Bradi1g56416.1	Bradi1g56458.1
–	6254296	6255252	Dup	4	No	No	Bradi1g09140.1	7	Bradi1g08870,1	Bradi1g08875.1
–	47990145	47991229	Dup	3	Yes	No	Bradi1g47460.1	3	Bradi1g49010.1	Bradi1g49015.1
–	50254598	50255514	Dup	2	No	No	Bradi3g16827.1	2	Bradi1g51595.1	Bradi1g51598.1
–	50738286	50739214	Dup	2	Yes	No	Bradi4g02308.1	3	Bradi1g52120.1	Bradi1g52140.2
–	31191057	31192043	Amb	2	Yes	No	Bradi3g28811.1	4	Bradi1g35496.1	Bradi1g35500.1
–	49059273	49060240	Amb	2	No	No	Bradi4g02308.1	3	Bradi1g50147.1	Bradi1g50147.2
–	42862180	42863160	Amb	1	No	No	Bradi3g06810.1	3	Bradi1g44440.1	Bradi1g44460.1
–	151228	152100	Amb	1	No	No	Bradi4g09271.1	4	Bradi1g00272.1	Bradi1g00278.1
–	32900117	32900976	Amb	1	No	No	Bradi1g10688.1	8	Bradi1g36971.1	Bradi1g36976.1

–	45623503	45624397	Dup	3	No	No	Bradi2g55860.1	4	Bradi1g46850.1	Bradi1g46860.1
–	58677033	58677883	Dup	4	Yes	No	Bradi1g69520.1	20	Bradi1g59300.1	Bradi1g59310.2
–	3820956	3821808	Dup	3	Yes	No	Bradi5g18560.1	6	Bradi1g05640.2	Bradi1g05650.1
–	46167944	46168776	Dup	3	No	No	Bradi1g43151.1	4	Bradi1g47407.2	Bradi1g47427.2
–	25276340	25277203	Amb	1	No	No	Bradi1g29693.1	1	Bradi1g29693.1	Bradi1g29697.1
–	62058436	62059207	Dup	2	No	No	Bradi1g41291.1	2	Bradi1g62547.1	Bradi1g62576.1
–	17062246	17062974	Amb	2	Yes	No	Bradi2g10656.1	3	Bradi1g21177.1	Bradi1g21190.1
–	55066109	55066910	Dup	2	No	No	Bradi1g28323.1	2	Bradi1g56095.1	Bradi1g56100.3
–	19804596	19805409	Dup	3	Yes	No	Bradi2g57850.1	3	Bradi1g24516.1	Bradi1g24528.1
–	35180715	35181428	Amb	1	Yes	No	Bradi1g58105.1	2	Bradi1g38730.1	Bradi1g38735.1
–	34023598	34024311	Amb	1	No	No	Bradi1g58105.1	2	Bradi1g37870.1	Bradi1g37876.1
–	64388717	64389453	Dup	2	Yes	No	Bradi1g21440.1	9	Bradi1g64957.1	Bradi1g64970.1
–	62890670	62891361	Amb	2	No	No	Bradi4g27965.1	2	Bradi1g63470.1	Bradi1g63480.1
–	34766131	34766848	Amb	2	No	No	Bradi2g44210.1	3	Bradi1g38350.1	Bradi1g38353.1
–	15657034	15657759	Amb	1	No	No	Bradi1g19580.1	4	Bradi1g19600.1	Bradi1g19607.1
–	41327863	41328507	Dup	3	Yes	No	Bradi3g22684.1	7	Bradi1g43380.1	Bradi1g43990.1
–	3822515	3823189	Dup	2	No	No	Bradi1g01670.1	11	Bradi1g05640.2	Bradi1g05650.1
–	24553681	24554343	Amb	2	No	No	Bradi4g06141.1	4	Bradi1g29025.1	Bradi1g29030.1
–	1323993	1324621	Amb	1	Yes	No	Bradi1g01950.1	1	Bradi1g01957.1	Bradi1g01965.1
–	72330203	72330813	Amb	1	Yes	No	Bradi1g75065.2	1	Bradi1g75060.1	Bradi1g75065.1
–	26068972	26069595	Amb	1	No	No	Bradi3g16827.1	2	Bradi1g30678.1	Bradi1g30690.1
–	26757753	26758379	Dup	2	Yes	No	Bradi4g22988.1	6	Bradi1g31193.4	Bradi1g31200.1
–	54223277	54223947	Amb	1	Yes	No	Bradi2g19723.1	1	Bradi1g55490.1	Bradi1g55500.1

(continued)

Table 3
(continued)

Pseudogene coordinates					Pseudogene features			Paternal locus		
Strand	Start	End	Type	Exons	Disablement	Poly-A	ID	Number exon	Locus_before	Locus_after
–	27521529	27522146	Amb	1	No	No	Bradi2g23660.1	1	Bradi1g32050.1	Bradi1g32060.1
–	57625094	57625762	Amb	1	Yes	No	Bradi1g28135.1	1	Bradi1g58545.1	Bradi1g58550.1
–	12729076	12729659	Amb	1	Yes	No	Bradi1g24126.1	1	Bradi1g15870.1	Bradi1g15790.1
–	61635309	61635895	Dup	2	No	No	Bradi2g36506.1	2	Bradi1g62057.2	Bradi1g62063.1
–	47699072	47699807	Dup	3	No	No	Bradi2g04706.1	3	Bradi1g48813.1	Bradi1g48817.1
–	22505013	22505669	Amb	1	No	No	Bradi4g13916.1	2	Bradi1g27388.1	Bradi1g27400.2
–	23440403	23441048	Dup	2	No	No	Bradi1g41720.1	2	Bradi1g28230.2	Bradi1g28260.2
–	4344947	4345528	Amb	1	No	No	Bradi1g28185.1	2	Bradi1g06470.1	Bradi1g06470.2
–	24099685	24100257	Amb	1	No	No	Bradi2g08380.1	1	Bradi1g12697.1	Bradi1g12701.1
+	26157800	26160537	Dup	6	Yes	Yes	Bradi5g09708.1	6	Bradi1g30750.1	Bradi1g30767.1
+	23557280	23559726	Dup	8	Yes	Yes	Bradi3g24672.1	8	Bradi1g28323.1	Bradi1g28326.1
+	29520135	29522422	Dup	6	Yes	Yes	Bradi4g22988.1	6	Bradi1g33850.1	Bradi1g33860.1
+	36286471	36288639	Dup	6	Yes	No	Bradi4g17405.1	6	Bradi1g39520.1	Bradi1g39620.1
+	24902945	24905024	Dup	4	Yes	No	Bradi1g43455.1	4	Bradi1g29340.1	Bradi1g29350.1
+	40472759	40474711	Dup	2	Yes	Yes	Bradi1g39454.1	3	Bradi1g42850.1	Bradi1g42860.1
+	23184495	23186333	Dup	9	Yes	Yes	Bradi1g28000.1	9	Bradi1g28000.1	Bradi1g28010.1
+	36287440	36289111	Dup	8	Yes	Yes	Bradi3g24672.1	8	Bradi1g39520.2	Bradi1g39620.1
+	9620465	9622087	Dup	4	Yes	Yes	Bradi1g14040.1	6	Bradi1g12470.2	Bradi1g12750.1
+	26910408	26912127	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g31326.1	Bradi1g31330.1

+	22914917	22916689	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g27780.1	Bradi1g27790.1
+	57237292	57239061	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g58105.1	Bradi1g58110.1
+	56289196	56290963	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g57271.1	Bradi1g57280.1
+	65138401	65140006	Dup	1	Yes	No	Bradi1g65990.1	1	Bradi1g65961.1	Bradi1g65964.1
+	7853574	7855339	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g10795.1	Bradi1g10800.1
+	32221866	32223629	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g36430.1	Bradi1g36441.1
+	8420495	8422257	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g11357.1	Bradi1g11365.1
+	28531162	28532797	Dup	4	Yes	Yes	Bradi1g05202.1	4	Bradi1g32890.1	Bradi1g32901.1
+	17393195	17394937	Dup	3	Yes	No	Bradi2g37365.1	3	Bradi1g21560.4	Bradi1g21570,1
+	43243537	43245173	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g44900.1	Bradi1g44920.1
+	35652036	35653766	Dup	3	Yes	Yes	Bradi3g23840.1	3	Bradi1g39056.1	Bradi1g39056.1
+	29441297	29443136	Dup	2	Yes	Yes	Bradi3g39278.1	3	Bradi1g33800.4	Bradi1g33803.1
+	32297063	32298801	Dup	3	Yes	Yes	Bradi1g21002.1	3	Bradi1g36505.1	Bradi1g36520.1
+	10869455	10871180	Dup	3	Yes	Yes	Bradi3g23840.1	3	Bradi1g13990.1	Bradi1g14000.1
+	18716041	18717602	Dup	9	Yes	Yes	Bradi1g21130.2	10	Bradi1g23293.1	Bradi1g23299.1
+	13059457	13061103	Dup	3	Yes	Yes	Bradi3g20965.1	4	Bradi1g16120.1	Bradi1g16140.1
+	5700790	5702344	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g08080.1	Bradi1g08094.2
+	8499041	8500566	Dup	2	Yes	Yes	Bradi1g39454.1	3	Bradi1g11430.1	Bradi1g11435.1
+	50028827	50030354	Dup	2	Yes	Yes	Bradi1g39454.1	3	Bradi1g51345.1	Bradi1g51350.1
+	24946382	24947915	Dup	3	Yes	Yes	Bradi3g20965.1	4	Bradi1g29356.1	Bradi1g29360.2
+	34730644	34732132	Dup	3	Yes	Yes	Bradi4g22988.1	6	Bradi1g38330.1	Bradi1g38340.1
+	4268444	4269990	Dup	1	Yes	Yes	Bradi4g20543.1	1	Bradi1g06372.1	Bradi1g06380.2
+	27119559	27121009	Dup	1	Yes	Yes	Bradi4g20543.1	1	Bradi1g31538.1	Bradi1g31547.1

(continued)

Table 3
(continued)

Pseudogene coordinates					Pseudogene features			Paternal locus		
Strand	Start	End	Type	Exons	Disablement	Poly-A	ID	Number exon	Locus_before	Locus_after
+	54198388	54199818	Dup	3	Yes	Yes	Bradi3g23840.1	3	Bradi1g55480.1	Bradi1g55485.1
+	36261140	36262419	Dup	3	Yes	Yes	Bradi1g05202.1	3	Bradi1g39490.1	Bradi1g39510.1
+	27835390	27836751	Amb	2	Yes	Yes	Bradi3g18720.2	3	Bradi1g32430.1	Bradi1g32437.1
+	36560167	36561406	Dup	3	Yes	Yes	Bradi2g37365.1	3	Bradi1g39720.1	Bradi1g39920.1
+	14049422	14050721	Dup	2	Yes	Yes	Bradi5g10831.1	2	Bradi1g17520.1	Bradi1g17526.1
+	100349	103382	Amb	1	Yes	Yes	Bradi4g09312.1	6	Bradi1g00237.1	Bradi1g00247.2
+	27927035	27928278	Amb	1	Yes	Yes	Bradi3g23840.1	3	Bradi1g32475.1	Bradi1g32480.1
+	59427320	59428564	Dup	2	Yes	Yes	Bradi3g28811.1	4	Bradi1g60010.1	Bradi1g60021.1
+	1181129	1182312	Amb	1	Yes	Yes	Bradi3g23840.1	3	Bradi1g01790.3	Bradi1g01795.1
+	41142262	41143466	Amb	1	Yes	Yes	Bradi1g43300.1	2	Bradi1g43271.1	Bradi1g43280.1
+	43019395	43020593	Dup	3	Yes	Yes	Bradi1g45085.1	5	Bradi1g44640.2	Bradi1g44640.4
+	89749	90873	Amb	1	Yes	Yes	Bradi4g09271.1	4	Bradi1g00232.1	Bradi1g00237.2
+	24946762	24947939	Dup	1	Yes	Yes	Bradi1g43455.1	4	Bradi1g29356.1	Bradi1g29360.2
+	11469405	11470497	Dup	3	Yes	Yes	Bradi4g35150.1	10	Bradi1g14517.1	Bradi1g14530.1
+	37033961	37035126	Dup	3	Yes	No	Bradi1g27841.1	5	Bradi1g40262.1	Bradi1g40290.1
+	71422822	71424010	Dup	3	Yes	Yes	Bradi1g73680.2	3	Bradi1g73690.1	Bradi1g73700.1
+	28531993	28533150	Dup	2	Yes	Yes	Bradi3g20965.1	4	Bradi1g32890.1	Bradi1g32901.1

References

1. International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768
2. Wicker T, Mayer KFX, Gundlach H et al (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. Plant Cell 23:1706–1718
3. Wang W, Zheng H, Fan C et al (2006) High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell 18:1791–1802
4. Guo X, Zhang Z, Gerstein MB et al (2009) Small RNAs originated from pseudogenes: cis- or trans-acting? PLoS Comput Biol 5: e1000449
5. Zhang Z, Carriero N, Zheng D et al (2006) PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics (Oxford) 22:1437–1439
6. Zhang ZD, Frankish A, Hunt T et al (2010) Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. Genome Biol 11: R26
7. Harrison PM (2014) Pseudogenes: functions and protocols. Springer, New York, NY
8. Thibaud-Nissen F, Ouyang S, Buell CR (2009) Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics 10:317
9. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410
10. Camiolo S, Porceddu A (2013) gff2sequence, a new user friendly tool for the generation of genomic sequences. BioData Mining 6:15
11. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
12. Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University
13. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988–995
14. Pearson WR, Wood T, Zhang Z et al (1997) Comparison of DNA sequences with protein sequences. Genomics 46:24–36

Chapter 13

TILLING in *Brachypodium distachyon*

Louise de Bang, Anna Maria Torp, and Søren K. Rasmussen

Abstract

TILLING is a low-cost screening method that allows for identification of mutations in a gene-of-interest within a range of few base pairs. TILLING can be applied to mutant populations or to plant collections of cultivars, landraces or crop wild relatives (Eco-TILLING). The method is based on the Cel1 enzyme cleavage of mismatches in PCR products amplified with labeled primers. The cleavage can be detected due to the labeled primers by different methods including capillary electrophoresis. Here, we introduce the development of the mutant population BRACHYLIFE and present a Brachypodium TILLING protocol based on fluorescing primers for PCR, enzymatic cleavage, and detection with Applied Biosystems 3130xl Genetic Analyzer.

Key words Sodium azide, Induced mutation, Cel1, Mutation detection

1 Introduction

Brachypodium has become a model plant for temperate grasses like ryegrass and the cereals such as wheat and barley. Since it was chosen as a model plant in the beginning of this century, a range of genomic tools and resources has been established and is still evolving [1, 2]. In particular collections of ecotypes and mutant populations offer possibilities of detecting alleles and new genes. In order to gain full value of chemically and physically induced mutations or even spontaneous mutations reverse genetics tools have been developed to identify sequence variants of a gene followed by evaluation of the phenotype. The high-throughput screening method TILLING (Targeting Induced Local Lesions IN Genomes) were introduced 15 years ago [3, 4] to identify mutations in genes-of-interest. After PCR amplification of the region to be searched, the detection of mutations is frequently based on either endonuclease cleavage of mismatches followed by electrophoresis or analysis of melting curves of double stranded PCR fragments by High Resolution Melt (HRM) analysis [5–9]. Analysis of melting curves by HRM is less laborious than the enzyme based method since it avoids any post-PCR handling steps. However, the length

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_13,

of PCR fragments that can be screened with HRM are less than 500 bp and typically 300 bp, where the length of the fragments with the enzyme based method is in the range of 800–1000 bp. Furthermore, the position of the mutation within the PCR fragment is not predicted with HRM [5–7]. The enzyme based method uses fluorescently labeled primers for PCR that allows identification of a mutation within a few nucleotides accuracy due to detection on either the LICOR gel system or by capillary electrophoresis [8, 9]. Using labeled primers for the PCR amplification of the gene-of-interest also allows automatic data collection and processing. It is however also possible to run the PCR without labeled primes followed by enzymatic cleavage and electrophoresis in agarose or polyacrylamide gels and staining with ethidium bromide [10, 11]. The development of a Brachypodium TILLING platform using sodium azide as mutagen and its use for screening by nuclease treatment and detection on the LICOR polyacrylamide equipment has been recently demonstrated [12]. Genomic resources and tools for Brachypodium still need to be extended in order to make it even more versatile as model plant. Here, we present the protocols used to develop a mutant population, including propagation and maintenance of lines as well as the TILLING protocol based on fluorescing primers for PCR, enzymatic cleavage, and detection with Applied Biosystems 3130*xl* Genetic Analyzer.

2 Materials

2.1 Mutagenesis

1. Sodium azide (S8032, Sigma-Aldrich) 1 M stock 1.625 g in 25 mL water.
2. Phosphate buffer 0.1 M pH 3.0 (13.6 g KH_2PO_4 in 1 L water adjust to pH 3.0).

2.2 DNA Extraction

1. BioSprint 96 DNA (Cat. No. 9000852, Qiagen).
2. BioSprint 96 DNA Plant Kit (Cat. No. 941557, Qiagen).
3. H_2O with Tween 20: 0.02% (v/v) Tween 20 (P1379, Sigma-Aldrich) in distilled water.
4. TE Elution buffer: 2 mM Tris–HCl pH 8.0 and 0.2 mM EDTA.
5. Isopropanol.
6. 96% ethanol.

2.3 Enzymatic Cleavage with Cel1

1. Cel1 enzyme (for the purification *see* [13]).
2. 10× Cel1 buffer (100 mM $MgSO_4$, 100 mM HEPES 100 mM KCl, 0.02% Triton X-100 (P1379, Sigma-Aldrich), and 2 μg/ml BSA).
3. 0.2 M EDTA.

2.4 Sephadex Purification

1. EMD Millipore MultiScreen™ column loader and scraper (MACL0SC03, Fisher scientific).
2. EMD Millipore MultiScreen™ filter plate 96-well (MSHVN 4510, Fisher Scientific).
3. Sephadex® G-50 Fine DNA Grade (17-0573-02, GE Healthcare).
4. Sterile H_2O.

2.5 Sample Preparation and Mutant Detection with Applied Biosystems 3130xl Genetic Analyzer

1. 0.1× TE buffer.
2. ROX size standard. In-house size standard with fragments ranging from 58 to 1259 bp.
3. Plate base 96-well (628-0155, Thermo Fisher).
4. Plate septa 96-well (435-933, Thermo Fisher).
5. Plate retainer 96-well (628-0160, Thermo Fisher).
6. Applied Biosystems 3130*xl* Genetic Analyzer (4399821, Thermo Fisher).
7. POP-7™ Polymer for 3130/3130*xl* Genetic Analyzers (4352759, Thermo Fisher).

2.6 Software

1. Program to analyze files from Applied Biosystems 3130*xl* Genetic Analyzer: GeneMarker® version 1.90 (SoftGenetics®).

2.7 Online Tools

1. Brachypodium sequence databases: Phytozome (http://phytozome.jgi.doe.gov/) and Gramene (http://gramene.org/).
2. Primer design program: Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/).
3. Mutation evaluation program: SIFT (http://sift.jcvi.org/).
4. Evaluation of potential effect of mutation on secondary structures: SAS (http://www.ebi.ac.uk/thornton-srv/databases/sas/) and SOPMA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html).

3 Methods

3.1 Development of Mutant Population

The development of the BRACHYLIFE *Brachypodium distachyon* Bd21-3 mutant population were based on the protocol developed for barley [14]. The mutagen treatment was carried out in sets of 4 × 100 seed for reasons of logistics and greenhouse capacity. The total number of seeds treated was approximately 3400 resulting in a mutant population of 3000 M2 plants. From 2504 of the M2 plants DNA was extracted and pooled for TILLING. The

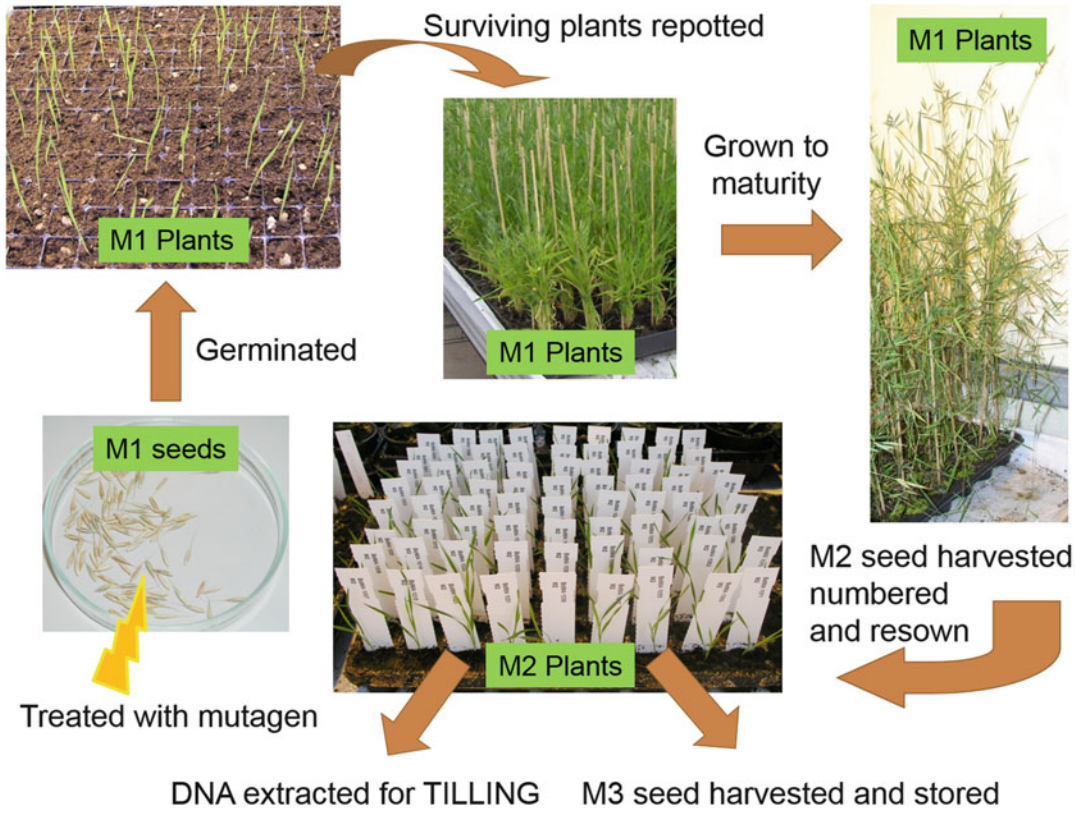

Fig. 1 Workflow in generation of the BRACHYLIFE mutant population. Seeds were incubated in the mutagen sodium azide, rinsed and germinated in small pots with soil. The Surviving M1 seedlings were repotted and grown to maturity. The M2 seeds were harvested from the mature M1 plants and numbered. M2 seeds were sown and fresh young leaf material was collected for DNA extraction for TILLING. The M2 plants were subsequently grown to maturity and the M3 seeds were harvested and stored at 4 ° C

concentration 20 mM sodium azide was chosen as mutagen strength based on an initial trial testing germination frequency of 10 mM, 15 mM, and 20 mM sodium azide.

Brachypodium distachyon Bd21-3 seeds were soaked in water for 2 h in a petri dish followed by treatment with 20 mM sodium azide in phosphate buffer pH 3 for another 2 h. The treatment was stopped by rinsing with water four times over a period of 1 h. Seeds (M1) were briefly dried and sown directly in soil. The seeds were germinated and surviving seedlings were repotted and grown for 2 weeks under short day condition, 8 h light (100 μE per m^2 per s) and 16 h dark at 20/20 °C followed by 16 h light (19 °C, 400 μE per m^2 per s) 8 h (17 °C, dark) until seed maturity. M2 seeds were harvested, dried and sown following the same growth conditions as for the M1 plants. Young leafs were harvested from the M2 plants for DNA extraction for TILLING. The M2 plants were grown to maturity and M3 seeds were harvested, bagged and stored at 4 ° C (*see* Fig. 1).

3.2 DNA Extraction and Normalization

A high-quality DNA extraction is preferred because production of plants and sampling of leafs are laborious with the great number of plants in a mutant population. The semiautomatic Qiagen BioSprint 96 DNA plant kit extraction method was chosen because of high quality and reduction in labor hours. Another DNA extraction method designed for TILLING called NEATILL (Nucleic acid Extraction from Arrayed Tissue for TILLING) is even less labor demanding and less expensive than the BioSprint method. NEATILL is based on direct multiplexing of plant tissue avoiding the extraction of DNA from all plants and the DNA normalization steps [15].

1. For DNA extraction the protocol *Purification of DNA from Plant Tissue Using the BioSprint 96* using fresh plant material (*see* **Note 1**) was followed with a minor change:

   Instead of applying one 3 mm tungsten-carbide bead after collection of plant material we applied two 4 mm glass beads before harvest. The glass beads are cheaper than tungsten-carbide beads and can therefore be discarded after use. For additional comments to the protocol *see* **Notes 2** and **3**.
2. Transfer extracted DNA from 96-well microplate MP to ThermoFisher 96-well plates (Cat 249944) and seal tight with the matching cap coat (Cat 276,002) for storage.
3. Measure DNA concentration. We used a NanoDrop™ 2000 spectrophotometer from Thermo Scientific with TE elution buffer (*see* Subheading 2.2, **item 4**) as reference.
4. Normalize all samples to the same concentration to ensure that each individual will be represented equally in the final DNA pool, we normalized all sample to 20 ng/μl (*see* **Note 4**).

3.3 DNA Pooling

When the DNA is normalized it has to be pooled for the TILLING screening. The number of DNA samples that can be combined in one pool depends on genome size and ploidy level of the plant species. Though classic TILLING uses eightfold pooling [3] we chose a tenfold one-way pooling strategy based on results from Lababidi et al. [16] which found this pooling depth to be appropriate for barley (*Hordeum vulgare* L.) in a similar setup. Brachypodium is diploid as barley but with a much smaller genome ~270 Mbp compared to ~3500 Mbp. The tenfold pooling allows a screen of 960 plants on one single plate.

1. Using an eight channeled pipet transfer one whole plate with individual plant DNA at a time to the pool plate as followed: A1-H1 from the individual plant DNA plate → A1-H1 in the pool plate, A2-H2 from the individual plant DNA plate → A2-H2 Pool plate and so on. First row in the pool plate will then represent first row of all ten individual plant DNA plates, the

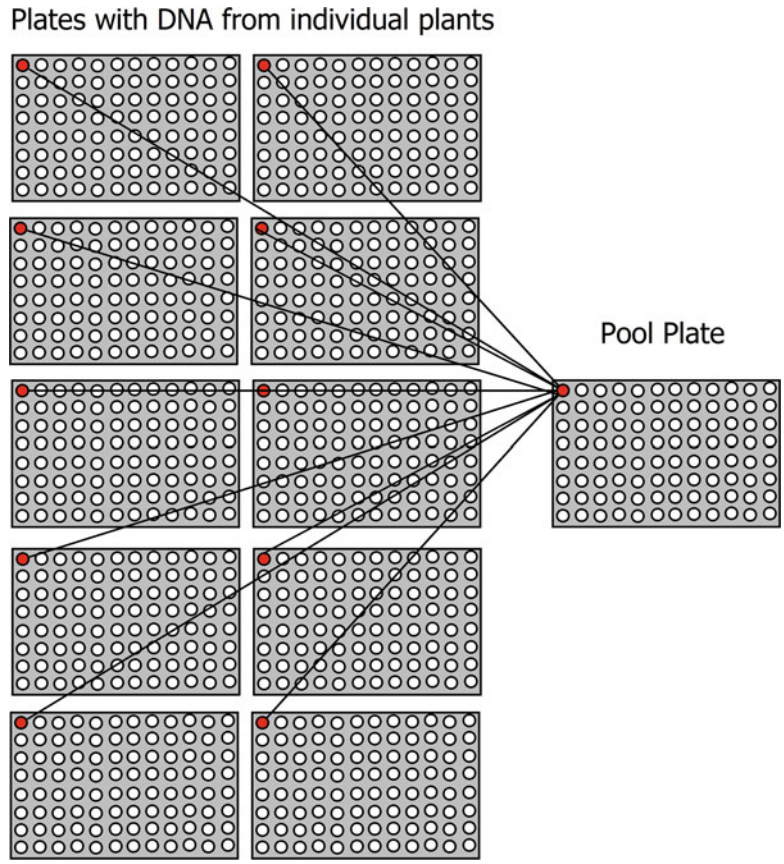

Fig. 2 Illustration of tenfold one-way pooling strategy. Normalized DNA from ten plates with DNA from individual plants is pooled in one single pool plate. The pool plate contains DNA from 960 individual plants that can be screened at once with the TILLING method

second row will represent all second rows of the ten individual plant DNA plates and so forth (*see* Fig. 2).

The advantages of the one-way pooling strategy selected is that rearrangement of the individual plant DNA is not needed and it reduces the number of plates to be screened initially compared to the two-way pooling strategy that are designed to screen all individual plants two times to start with. On the other hand when a putative mutation is discovered in the initial screen, the one-way pooling strategy needs a secondary screen of individual plant DNA added wild type DNA, which is not necessary in the two-way pooling strategy [8].

3.4 Target Gene

The target genes for TILLING are usually genes with a known function that is expected to cause a visual or measurable phenotype. An example is the BRI1 gene. BRI1 is the receptor of the growth hormone brassinosteroid and several studies of *bri1* mutants reveal

severe dwarfed phenotypes; *Arabidopsis thaliana* [17], rice (*Oryza sativa*) [18], and maize (*Zea mays*) [19].

After a gene has been selected the next step is to design primers. Specificity of primers is very important in TILLING, so a bioinformatics study should be carried out in order to reveal potential homolog genes. The availability of the complete Brachypodium genomic sequence facilitates easy gene specific primer design and if homolog genes appear, the sequence information can be used to design primers in the variable areas of the genes. Sequences of Brachypodium can be found on several public databases among others Phytozome (http://phytozome.jgi.doe.gov/) and Gramene (http://gramene.org/).

3.5 PCR Primer Design

To cover a whole gene several primer pairs are usually needed, but in order to reduce the number of target fragments the primers should be placed wisely to cover as much of high priority area such as exon and promoter regions and leave out introns. In addition primers can be directed toward regions of the genes encoding highly conserved regions of the protein (*see* **Note 5**). We have used the online program CODDLE to detect areas in genes-of-interest with highest likelihood of mutagen susceptibility as a tool for primer design. Unfortunately this program is not available anymore and to our knowledge no other programs are at the moment available that can replace this function.

The optimum size of the target fragment is between 800–1100 bp and primers should be selected with melting temperatures of approximately 70 °C (range from 67 to 73 °C) and an optimum length of around 27 bp (range from 20 to 30 bp) [8]. Online primer design programs are available of which we have used Primer3 (http://bioinfo.ut.ee/primer3-0.4.0/primer3/) (*see* **Note 6**).

3.6 Fluorescing Primers

For capillary electrophoresis on the ABI 3130*xl* Genetic Analyzer forward primers are normally labeled with 6FAM fluorescence dye that have excitation at 485 nm and emission at 520 nm, thus emitting in the blue part of visible spectrum. The reverse primers are usually labeled with VIC that excite at 538 and emits in the yellow/green part of the visible spectrum at 554 nm (*see* **Note 7**).

3.7 Mutation Screening

3.7.1 PCR Amplification and Heteroduplex Formation

A PCR-mixture for 96 samples is set up with the primer mix (*see* **Note 7**) and the proportions of dNTP, H_2O and specific buffer and polymerase that meets the requirements of the specific polymerase (*see* **Note 6**) selected. We use 8.5 µl PCR mix to 1.5 µl DNA. Transfer PCR mix with an automatic pipet to all wells in a 96-well PCR plate. Then transfer, to the PCR plate, DNA samples from a complete DNA pool plate with an 8 channeled pipet while keeping track of the DNA samples. Always place the DNA sample in the

Table 1
Example of program for PCR amplification and heteroduplex formation with the annealing temperature 76 °C

Step 1		Step 2		Step 3		Number Cycles
Temp.	Time	Temp.	Time	Temp.	Time	
Amplification						
94 °C	2 min					
94 °C	30 s	76 °C[a]	30 s	72 °C	1 min	8
94 °C	30 s	68 °C	30 s	72 °C	1 min	45
Heteroduplex formation						
72 °C	5 min					
99 °C	10 min					
70 °C[b]	30 s					70
4 °C	∞					

[a]Decrease 1 °C per cycle
[b]Decrease 0.3 °C per cycle

PCR plate in the exact same well as in the pool plate. Seal the plate with a PCR seal and spin down briefly (*see* **Note 8**).

PCR amplification of the gene of interest and heteroduplex formation of the PCR product can be carried out in the same program on the PCR machine (*see* Table 1).

The PCR program should meet the requirements of the primers optimal annealing temperature and the extension time for the specific polymerase.

3.7.2 Cel1 Digestion

To each 10 μl PCR sample add 20 μl Cel1 mix containing 16.9 μl ddH_2O, 3.0 μl 10× Cel1 buffer and 0.1 μl Cel1 enzymes while the samples are kept on ice.

Use a PCR machine to incubate samples at 45 °C for 60 min. Transfer samples to ice and stop the Cel1 digestion by adding 5 μl 0.2 M EDTA to each sample.

3.7.3 Purification of Samples with Sephadex Column

It is important to remove buffer salts from the samples because salt anions will compete with the negatively charged DNA during capillary electro kinetic injection on the ABI 3130*xl* Genetic Analyzer.

Any purification method that removes salt and buffer and reduces sample volume can be applied [8]. The following method uses the Sephadex G-50 column.

Place Sephadex G-50 powder on the Millipore column loader. With the supplied scraper distribute the powder to all wells and

remove excess powder. Place the Millipore 96-well filter plate on top of the Sephadex containing Millipore column loader. While keeping the two plates tightly together turn them around and carefully knock them to ensure that all powder enters the Millipore 96-well filter plate. With an eight-channeled pipet add 300 µl sterile ddH_2O to each well, cover with Parafilm and a plastic bag and leave the plate in the fridge for at least 2 h (*see* **Note 9**).

Place the Sephadex containing Millipore 96-well filter plate on top of a 96-well waste plate and centrifuge at 2000 rpm (805 × *g*) for 5 min. Discard water from the waste plate and pipet 150 µl sterile H_2O to the wells of the Sephadex containing plate. Again centrifuge at 2000 rpm (805 × *g*) for 5 min and discard water and place the Sephadex containing plate on top of a new 96-well PCR elution plate. Pipet the Cel1 digested samples to the middle of the filters in the wells and leave for 3 min. at room temperature (*see* **Note 10**).

Centrifuge at 2000 rpm (805 × *g*) for 5 min. (*see* **Note 11**). The purified samples can at this point be sealed and stored at −20 °C.

*3.7.4 Mutation Detection with Applied Biosystems 3130**xl** Genetic Analyzer*

It is almost always necessary to dilute the purified PCR product before analyzing on the ABI 3130*xl* Genetic Analyzer. We found that a dilution of approximately 1:6 were optimal.

In a microtiter plate pipet 12 µl 0.1% TE buffer to all wells in the plate, then with an eight channeled pipet transfer 2–3 µl PCR product to each well and ad 0.5 µl ROX standard. Keep the same arrangement of samples in the new plate. Spin briefly and use a PCR machine to denature all samples at 94 °C for 2 min.

Mount the microtiter plate in the Plate base 96-well, add the Plate base 96-well to cover the microtiter plate, and at last cover with the Plate retainer 96-well.

We use an Applied Biosystems 3130*xl* DNA genetic analyzer with a 36 cm capillary array and POP-7 polymer. The images are analyzed for presence of mutations using the software GeneMarker® version 1.90. A sample containing a mutation will have two unique peaks/bands one green and one blue representing each end label of the PCR product. The length of the two unique peaks/bands should always add up to the total length of the initial PCR fragment. Use the added ROX standard sizes to estimate lengths of products (*see* Fig. 3).

3.8 Mutation Identification and Evaluation

When a pool sample are identified with a putative mutation, the individual plant containing the putative mutation are found by analyzing DNA from all individual plants in the specific pool following the same procedure as for the pooled DNA (Subheading 3.7, **steps 1–4**). When screening single plant DNA, wild type DNA should always be added because the TILLING procedure recognizes mismatches of the paired strands. If wild type DNA is not applied a homozygous mutation will not be discovered by the

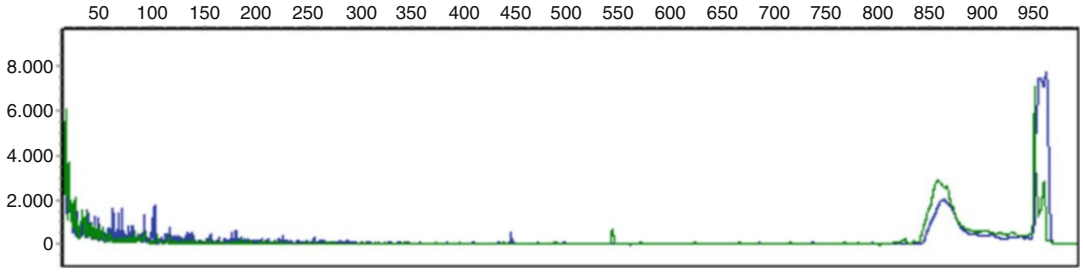

Fig. 3 Allele report with fluorescing intensity along the length of the PCR product from GeneMarker V1.90 SoftGenetics showing a mutant allele. Fluorescing intensity from both the blue and the green dye (representing PCR amplification from both ends of the PCR fragment) has been overlaid, showing a clear mutant allele. The *blue* and the *green* peak represent each end of the PCR product that has been cut in two by the enzyme CEL1 that recognizes mismatches in heteroduplexes. The mismatch appears because a mutant allele is paired up with a wild type allele during the heteroduplex formation. The first part of the mutant fragment (the *blue* peak) is around 446 bp long and the last part of the mutant fragment (the *green* peak) is around 544 bp long. Both fragments add approximately up to the full length size of the PCR product which is 981 bp

method because it does not form mismatches during the heteroduplex formation step (Subheading 3.7, **step 1**).

Once the putative mutation are tracked down to DNA from one single plant it can be confirmed by sequencing. Amplify the mutation containing fragment of the individual plant with PCR and purify it for sequencing. We use E.Z.N.A.® Gel Extraction Kit and the sequencing service GATC Biotech (http://www.gatc-biotech.com) but any purification and sequencing service can be applied. Using any sequence analysis software the mutation can be identified and characterized by the following parameters:

Mutation site: exon, intron, splice site or promoter region.

Mutation type: Base pair change or an insertion/deletion (this is usually dependent on the mutagen method).

Mutation effect: silent, missense, nonsense, or a frameshift.

Several online tools to predict and evaluate the effect of mutations are available. A program described by Ng and Henikoff [20] called SIFT (Sorting Intolerant From Tolerant; http://sift.jcvi.org/) can predict the effect/severity of a mutation by evaluating amino acid sequences of homolog genes. The putative effect of a mutation on the secondary structure of the protein can be evaluated by online programs such as SAS (Sequence Annotated by Structure; http://www.ebi.ac.uk/thornton-srv/databases/sas/) and SOPMA (Self-Optimized Method for secondary structure prediction with Alignment; http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html).

Once a mutation is identified and it is expected to have influence on gene function, seeds from the specific M2 plant are sown out and the offspring are analyzed for heterozygosity or

homozygosity of the mutation. Homozygous plants are raised to increase seed lot of the mutant and backcrossing to wild type can begin in order to remove background mutations. Along this work analysis of the expected phenotype of the mutant can be carried out to get an impression of the mutations effect (*see* **Note 12**).

3.9 Example

This TILLING method was used to screen the above described Brachypodium mutant population in the BRI1 gene (Bradi2g48280.1). A mutation was found that caused a G to A transition. This base change caused a change in amino acid 815 in the protein from Glycine (GGG) to Glutamic acid (GAG). Effects of changes in this site was analyzed with the public internet program SIFT. The analysis was performed with 100 homolog sequences found in a NCBI BLink BLAST with the GI number of Bradi2g48280.1 (XP_003569690) and standard settings in the SIFT BLink function. The program predicted that a change from Glycine to Glutamic acid in this position was intolerated with a SIFT Score of 0.00 (A SIFT score range from 0 to 1. A number <0.05 = in tolerated and a number >0.05 = tolerated). Even before backcrossing, the phenotype that follows the segregation of the mutation is very clear (*see* Fig. 4).

3.10 Mutation Frequency

Based on the amount of sequences screened with TILLING the mutation frequency of the BRACHYLIFE population is found to be 0.7 mutation pr 10^6 bp or one mutation pr. 1443 bp (*see*

Fig. 4 *bri1* mutant phenotype. The *bri1* mutant (*left*) is compared to its sibling without the *bri1* mutation (*right*)

Table 2
Mutation frequency in the BRACHYLIFE mutant population

	Population size	Total amount of screened, bp	GC content	Mutations identified	Mutation frequency
BRACHYLIFE	2504	17.3×10^6	51	12	1/1443

Mutation frequency is based on total amount of base pairs screened and mutations identified. The GC content is not taken into account in the calculation

Table 2). Another published Brachypodium mutant population BRACHYTIL has a mutation rate on 1/396 mutation/bp [12]. Some of the differences between mutation frequencies can lie within the total amount of sequence that has been screened, which is three times more in the BRACHYTILL population. GC content can also have an influence on mutation frequency.

4 Notes

1. We collected 2× approximately 5 cm young leaf tissue. In case where DNA was not extracted right away the leaf material was kept at −80 °C.
2. **Step 1** in purification protocol: When preparing the plates for the BioSprint, pipette the RPW buffer first. The buffer is kept in the fridge to keep the RNaseA working. However, the buffer is not working well when cold. Pipetting this buffer first ensures that the buffer reaches room temperature before the run.
3. **Step 4** in purification protocol: Apply MagAttract solution in the middle of the well, as the beads tends to stick to the sides of the well.
4. To protect the DNA for long term storage, keep the extracted DNA in freezer (around −20 °C). The normalized DNA and the pools can be stored at 4 °C for daily use.
5. Though the primer pairs should be designed to cover high priority areas, they should at the same time be placed with a certain margin of 50–100 bp to the target area. This is because short fragments are difficult to detect on the final gel or image.
6. Test and optimize the primer efficiency in regard to both temperature and polymerase of unlabeled primers before ordering the more expensive labeled primers, the choice of polymerase may depend on the GC content of the PCR fragment. However, be aware that labeled primers frequently works better at a slightly lower annealing temperature compared to unlabeled primers, usually 2 °C lower.

7. To optimize PCR output, always use the labeled primers in a mixture with unlabeled primers. The fluorescing labels hampers the primers' ability to reach the genomic DNA. Using a primer mix, unlabeled primers copies the genomic sequence and releases smaller strands that are easier for the labeled primers to process. Mixture of primers for Applied Biosystems 3130*xl* Genetic Analyzer: From primer stocks (100 μM) we mix 3.0 μl of each forward and reverse primer and 2 μl of each VIC and 6-FAM labeled primer in a total of 200 μl. Always make fresh aliquots of primer mixes and keep the stocks of fluorescence labeled primers protected from light.
8. Always keep the workspace for PCR setup and all the post-PCR handling steps separated to minimize the risk of PCR carryover contamination. Due to the extensive handling of PCR products in the Cel1 based TILLING protocol, the TILLING method is sensitive to contamination.
9. To save time the preparation of the Sephadex filter can be carried out before the setup of the PCR reaction (Subheading 3.7, **step 1**) or earlier. The hydrated Sephadex filters can be kept at 4 °C for a week when sealed with Parafilm and covered with a plastic bag.
10. Make sure to keep track of the samples. Always place the Sephadex filter plate the same way as the elution plate underneath (A1 on top of A1) and pipet the Cel1 digested samples so that they end up in the exact same well in the elution plate.
11. The Milipore 96-well separation plate can be reused for around ten times. Discard the Sephadex® filters and carefully clean the plate with ion-exchanged water. Leave the plate for drying, and do not reuse before it is completely dry.
12. When analyzing effect of the mutation before backcrossing siblings to the mutant without the specific mutation is a better reference than the wild type because the background genome will be more similar.

References

1. Catalan P, Chalhoub B, Chochois V, Garvin DF, Hasterok R, Manzaneda AJ et al (2014) Update on the genomics and basic biology of Brachypodium: international Brachypodium initiative (IBI). Trends Plant Sci 19 (7):414–418
2. Bragg JN, Wu J, Gordon SP, Guttman ME, Thilmony R, Lazo GR et al (2012) Generation and characterization of the Western Regional Research Center Brachypodium T-DNA insertional mutant collection. PLoS One 7(9): e41916
3. Colbert T, Till BJ, Tompa R, Reynolds S, Steine MN, Yeung AT et al (2001) High-throughput screening for induced point mutations. Plant Physiol 126(2):480–484
4. McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeting induced local lesions in genomes (TILLING) for plant functional genomics. Plant Physiol 123(2):439–442
5. Gady ALF, Hermans FWK, Van de Wal MHBJ, van Loo EN, Visser RGF, Bachem CWB (2009) Implementation of two high throughput techniques in a novel application: detecting

point mutations in large EMS mutated plant populations. Plant Methods 5:13

6. Dong CM, Vincent K, Sharp P (2009) Simultaneous mutation detection of three homoeologous genes in wheat by high resolution melting analysis and mutation surveyor (R). BMC Plant Biol 9:143
7. Botticella E, Sestili F, Hernandez-Lopez A, Phillips A, Lafiandra D (2011) High resolution melting analysis for the detection of EMS induced mutations in wheat SbeIIa genes. BMC Plant Biol 11:156
8. Till BJ, Zerr T, Comai L, Henikoff S (2006) A protocol for TILLING and ecotilling in plants and animals. Nat Protoc 1(5):2465–2477
9. Le Signor C, Savois V, Aubert G, Verdier J, Nicolas M, Pagny G et al (2009) Optimizing TILLING populations for reverse genetics in *Medicago truncatula*. Plant Biotechnol J 7 (5):430–441
10. Uauy C, Paraiso F, Colasuonno P, Tran RK, Tsai H, Berardi S et al (2009) A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat. BMC Plant Biol 9:115
11. Raghavan C, Naredo MEB, Wang HH, Atienza G, Liu B, Qiu FL et al (2007) Rapid method for detecting SNPs on agarose gels and its application in candidate gene mapping. Mol Breed 19(2):87–101
12. Dalmais M, Antelme S, Ho-Yue-Kuang S, Wang Y, Darracq O, d'Yvoire MB et al (2013) A TILLING platform for functional genomics in *Brachypodium distachyon*. PLoS One 8(6): e65503
13. Mejlhede N, Kyjovska Z, Backes G, Burhenne K, Rasmussen SK, Jahoor A (2006) EcoTILLING for the identification of allelic variation in the podery mildew resistance genes mlo and Mla of barley. Plant Breed 125:461
14. Anonymous 1977 Manual on mutation breeding, 2nd edn. Agency IAE (ed) Joint FAO/ IAEA Division of Atomic Energy in Food and Agriculture, Vienna
15. Sreelakshmi Y, Gupta S, Bodanapu R, Chauhan VS, Hanjabam M, Thomas S et al (2010) NEATTILL: a simplified procedure for nucleic acid extraction from arrayed tissue for TILLING and other high-throughput reverse genetic applications. Plant Methods 6(1):3
16. Lababidi S, Mejlhede N, Rasmussen SK, Backes G, Al-Said W, Baum M et al (2009) Identification of barley mutants in the cultivar 'Lux' at the Dhn loci through TILLING. Plant Breed 128(4):332–336
17. Clouse SD, Langford M, McMorris TC (1996) A brassinosteroid-insensitive mutant in *Arabidopsis thaliana* exhibits multiple defects in growth and development. Plant Physiol 111 (3):671–678
18. Morinaka Y, Sakamoto T, Inukai Y, Agetsuma M, Kitano H, Ashikari M et al (2006) Morphological alteration caused by brassinosteroid insensitivity increases the biomass and grain production of rice. Plant Physiol 141 (3):924–931
19. Kir G, Ye H, Nelissen H, Neelakandan AK, Kusnandar AS, Luo A et al (2015) RNA interference knockdown of BRASSINOSTEROID INSENSITIVE1 in maize reveals novel functions for brassinosteroid signaling in controlling plant architecture. Plant Physiol 169(1):826–839
20. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814

Chapter 14

Method for the Large-Scale Identification of phasiRNAs in *Brachypodium distachyon*

Kun Yang, Xiaopeng Wen, and Gaurav Sablok

Abstract

Postranscriptional regulation has been widely shown to be regulated by several classes of small non-coding RNAs; most abundantly, microRNAs, which have been shown to be the first dominant class and has been widely characterized as post-transcriptional regulators. In addition to microRNAs, triggered by miRNAs, transcripts called as *PHAS* (or *TAS*) generate abundant class of small RNAs in 21-nt manner, which is a pattern formed by DICER-LIKE 4 (DCL4) processing. Although *PHAS* can be identified by aligning transcripts to reported *PHAS* in other species, the most sensitive and accurate way to discovery them is by mapping of the smallRNAs taking into account the transcript coordinates. Here, we describe a workflow that can be used for the identification *PHAS* and corresponding phasiRNAs in *Brachypodium distachyon* using publically availabe smallRNAs datasets.

Key words *PHAS*, phasiRNA, Hyper geometric distribution, Phasing score, Bioinformatics

1 Introduction

Regulatory RNAs play an important role in fine tuning the post-transcriptional landscape, among which small noncoding RNAs, which are generated in intervals of 21- or 24-nt, which are often described as phases has recently been shown to have regulatory roles. Based on the phased origin of these RNAs, they have been primarily called as phasiRNAs and have been shown to play an important roles in reproductive development and regulating their targets in *trans*, thus called *trans*-acting siRNAs [1]. In particular, 21- and 24-nt reproductive phasiRNAs, triggered by miR2118 and miR2275, respectively, have been shown to critical regulators of the gametogenesis in maize [2]. Recent reports indicate regulatory roles of phasiRNAs, which are associated with photoperiod-sensitive genic male sterility 1 (Pms1) in regulating the photoperiod-sensitive male sterility in rice [3]. Taking into

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_14,

account the increasing role of the phasiRNAs as master regulators, it can be presumed that this emerging class of siRNAs can reveal new regulatory roles to enhance our understanding of post-transcriptional regulation of plant stress and development.

The origin and distribution of these phasiRNAs and locus generating these phasiRNAs (PHAS Loci) have been seen randomly distributed through the genome. Although the origin of phasiRNAs is still a debated issue with majority of the reports concluding the biogenesis through RDR6-dependent small RNA biogenesis, their roles in regulating the development and meiotic changes have been widely demonstrated [4–6]. Briefly, transcripts are cleaved by miRNA triggers. In *Arabidopsis*, two miRNA triggers for one transcript are needed to generate phasiRNAs whereas some loci identified from legume species require only one trigger miRNA [7]. After cleavage by trigger miRNAs, reverse complementary strand is synthesized by an endogenous RNA-dependent RNA polymerase using one of the products as template [8, 9]. Following, perfectly paired double-strand RNAs are recognized by DCL and processed in a phased manner to generate many small RNA duplex in same length [10]. Similar with miRNA duplexes, these ones can be recruited by members of ARGONAUTE protein and direct the RISC to target mRNAs [10]. The origin and distribution of these phasiRNAs and locus generating these phasiRNAs (PHAS Loci) have been seen randomly distributed throughout the genome [11]. In *Arabidopsis*, eight PHAS loci were previously identified, while hundreds of PHAS loci were predicted in other plant species [11].

Here, we describe a methodology, which can predict *PHAS* loci through mapping smallRNA-Seq datasets to the genome or other substituted sequence sets. This methodology takes many important features that *PHAS* loci have into consideration and can be used for identification of *PHAS* loci processing phasiRNAs other than 21-nt.

2 Materials

1. A FASTA file containing collapsed small RNAs with the header reflecting the following information:

   >SRR_1_x77275.

   TCGGACCAGGCTTCATTCCCC.

2. A FASTA file consists of nuclear genome or transcriptome of the same species (*see* **Note 1**).
3. A computer with Perl, R, bowtie, and bowtie-build available (*see* **Note 2**).

3 Methods

3.1 Building Genome Indices

This step can be simply achieved by running bowtie-build with the default parameters.

PSEUDOCODE: $bowtie-build -f Brachypodium_genome.fa Brachypodium_genome.

3.2 Mapping Small RNAs to Genome Indices

Coordinates of the 5′ end in the genome is used as the origin position of the mapped small RNAs. Small RNAs mapped in the antisense strand will include 2-nt positive offset to mimic the 3′ end overhanging.

PSEUDOCODE: $bowtie -f -m 6 -v 0 -a -p 4 Brachypodium_genome smallRNA.fa smallRNA.bwt

(a) For each valid mapping, its record should cover the following information at least:

- Header of the small RNA.
- Read count of the small RNA.
- Sequence of small RNA or the alignment.
- Header of the genome sequence.
- Strand of mapped.
- Five prime coordinate of the small RNA in the mapped genome sequence.
- Three prime coordinate of the small RNA or the length of the mapped region.

Multi-mapping reads mapping to multiple loci; mostly, those whose sequence is simple and are mapped to repeat regions should be discarded to decrease the background noise.

3.3 Extracting Small RNA Hotspots from Genome

This methodology uses a sliding window and walks through the reference sequence to find the positive segments generating small RNAs in phased manner. Sliding window walking through the entire genome is time consuming and also memory-consuming. To improve efficiency, small RNA hotspots should be extracted, and the walking will be carried out in these hotspots (*see* **Note 3**).

PSEUDOCODE:

(a) For each hotspot, following information should be recorded:

- Headers of genome sequence.
- Start coordinates in the genome sequence.
- End coordinates in the genome sequence.

3.4 Identifying Positive Sliding Windows

Workflow under this subtitle assumes to identify regions generating small RNA in 21-nt manner. To identify regions generating small RNA in other manners, parameters should be changed accordingly.

This step uses a 189-nt (9-cycle) sliding window walking in the hotspots with 1-nt each step.

1. Assigning each nucleotide coordinate within the window a *PHAS* register which is 21-nt is used here. This register groups small RNAs which are generated in the same manner. For example, the *PHAS* register of coordinate 213 = 213% 21 = 3 and coordinate 234 = 234% 21 = 3. The above calculation means coordinates 213 and coordinate 234 are both generated in same 21-nt manner. The register of 220 = 220% 21 = 10, which means coordinate 220 is in a different 21 manner. For each window, the *PHAS* register 1st (*see* **Note 4**) should be the one generating phasiRNAs.
2. Filtering windows showing low possibility to generate small RNA in 21-nt manner.
   PSEUDOCODE:
   (a) For each window:
   - Calculating number of unique small RNAs.
   - Calculating number of unique 21-nt small RNAs.
   - Calculating number of *PHAS* register 1st occupied by 21-nt small RNAs;
   - Discarding those with only a few small RNAs mapped, those with low ratio of uniquely mapped 21-nt small RNAs, and those with low proportion of *PHAS* registers occupied by 21-nt small RNAs.
3. Filtering windows with high *P*-value:
   PSEUDOCODE:
   - Calculating number of unique 21-nt small RNAs.
   - Calculating number of *PHAS* register 1st occupied by 21-nt small RNAs.
   - Calculating *P*-value using following formula:

$$P\text{-value}: p(k) = \sum_{X=k}^{m} \left[\frac{\binom{(l-1)m}{n-k}\binom{m}{k}}{\binom{lm}{n}} \right],$$

   where l = the manner of the phase (21-nt), m = number of phasing cycle (9-cycle), n = number of unique small RNAs, k = number of *PHAS* register 1st occupied by 21-nt small RNAs.
   - Discarding sliding windows whose *P*-value is high.

   (b) For each positive sliding window, recording the following information:

- Header of the genome sequence.
- Start coordinate in the genome.
- End coordinate in the genome.

3.5 Calculating Phasing Score

1. Merging positive sliding windows whose *PHAS* register 1st are the same as *PHAS* candidates.

   PSEUDOCODE:

   (a) Grouping positive sliding windows by their *PHAS* register 1st.

   (b) Ordering positive sliding windows in the same group by their start coordinates.

   (c) Checking start and end coordinates of the adjacent positive sliding windows to see if they have overlaps.

   (d) Extending the positive window by taking the most left and right coordinates of overlapped adjacent windows until they don't have overlaps anymore.

2. Filtering *PHAS* candidates with highest phasing score.

   PSEUDOCODE:

   (a) For each *PHAS* candidate:

   - Calculating number of 21-nt small RNAs.
   - Calculating number of *PHAS* register 1st occupied by 21-nt small RNAs.
   - Calculating number of 21-nt small RNAs occupying *PHAS* register 1st.
   - Calculating phasing score of each coordinate with following formula.
   - For *PHAS* candidates, each coordinate is regarded as the 95th position, which is the middle position of a sliding window. Its phasing score is calculated with the following formula:

$$\text{Phasingscore} = \ln\left[\left(1 + 9 \times \left(\frac{\sum_{i=1}^{9} P_i}{1 + \sum U}\right)\right)^{(n-2)}\right], n > 3,$$

     where n = number of PHASE register 1st occupied by 21-nt small RNAs, P = number of 21-nt small RNAs occupying PHASE register 1st and U = number of 21-nt small RNAs occupying other PHASE registers.

   - Discarding *PHAS* candidates whose highest phasing scores are small.

   An example of identified *PHAS* loci is shown in Table 1.

Table 1
Identified 21-nt *PHAS* loci in *Brachypodium*

Chr	Register 1st	Start	End	SRR1174000	SRR1174001	SRR1174002	SRR1174003	SRR1174004	SRR1174005	SRR1174006	SRR1174007
Bd1	18	63,642,240	63,642,680	+	+	+	+	+	+	+	+
Bd1	19	63,642,241	63,642,513					+			
Bd2	7	2,055,193	2,055,381				+				
Bd2	17	38,847,623	38,847,874		+		+	+			+
Bd3	5	2,605,580	2,605,810			+					
Bd3	15	49,960,569	49,961,156	+	+	+	+	+	+	+	+
Bd4	3	9,830,628	9,830,984	+	+	+	+	+	+	+	+
Bd4	6	9,674,643	9,674,831							+	
Bd4	16	9,891,121	9,891,477			+	+	+	+	+	
Bd4	19	9,654,454	9,654,873			+	+	+	+	+	
Bd4	19	9,840,892	9,841,143				+			+	
Bd5	7	18,150,328	18,150,747		+	+	+	+	+	+	+

The gray cells indicate locus is sensitive to this methodology in corresponding dataset

3.6 Extracting and Quantifying phasiRNAs from Identified PHAS

As the 5′ coordinates of all phasiRNAs generated by identified *PHAS* are at *PHAS* register 1st, they can be simply extracted from column3 and column4 of Subheading 3.2 result by column5 and column6.

4 Notes

1. For species whose assembled genome is not available, transcriptome, Expressed Sequence Tags (ESTs) sequences or subsets of the genome in the form of BAC or FOSMID libraries can be used as an alternative.
2. The Perl and R are not mandatory. User can choose other programming language and statistic tool based on which they are familiar. The same with bowtie and bowtie-build.
3. This step is only needed when using assembled genome as reference. Each transcript, ESTs sequence or shotgun sequence of genome can be treated as a hotspot. When you are trying to extract the hotspot from the genome, the most important thing is to know for how long the distance of two mapped 5′ coordinates should be regarded as from different hotspots. Normally, different criteria to extract the hotspots will only affect the efficiency of this methodology but not the final result.
4. PHASE register 1st and PHASE register 11st represent the PHASE register of the first and eleventh coordinate. For example, if the positive window lies between the 233 and 421 coordinate in the chromosome 3 of the genome, then the 1st PHASE register is 2 = 233% 21 here and the coordinates in this register should be 233, 254, 275, ..., 401 and the PHASE register 11st is 12 = 243% 21 here and the coordinates in this register should be 243, 264, 285, ..., 411.

References

1. Axtell MJ (2013) Classification and comparison of small RNAs from plants. Annu Rev Plant Biol 64:137–159
2. Zhai J, Zhang H, Arikit S, Huang K, Nan GL, Walbot V, Meyers BC (2015) Spatiotemporally dynamic, cell-type-dependent premeiotic and meiotic phasiRNAs in maize anthers. Proc Natl Acad Sci U S A 112:3146–3151
3. Fan Y, Yang J, Mathioni SM, Yu J, Shen J, Yang X, Wang L, Zhang Q, Cai Z, Xu C, Li X, Xiao J, Meyers BC, Zhang Q (2016) PMS1T, producing phased small-interfering RNAs, regulates photoperiod-sensitive male sterility in rice. Proc Natl Acad Sci U S A 113:15144–15149
4. Johnson C et al (2009) Clusters and superclusters of phased small RNAs in the de-veloping inflorescence of rice. Genome Res 19 (8):1429–1440
5. Komiya R et al (2014) Rice germline-specific argonaute MEL1 protein binds to phasiRNAs generated from more than 700 lincRNAs. Plant J 78(3):385–397
6. Dukowic-Schulze S, Sundararajan A, Ramaraj T, Kianian S, Pawlowski WP, Mudge J, Chen C

(2016) Novel meiotic miRNAs and indications for a role of PhasiRNAs in meiosis. Front Plant Sci 7:762

7. Zhai J, Jeong DH, De Paoli E, Park S, Rosen BD, Li Y et al (2011) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes Dev 25(23):2540–2553
8. Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS (2004) SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. Genes Dev 18(19):2368–2379
9. Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC et al (2004) Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Mol Cell 16(1):69–79
10. Yoshikawa M, Peragine A, Park MY, Poethig RS (2005) A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Dev 19(18):2164–2175
11. Zheng Y, Wang Y, Wu J, Ding B, Fei Z (2015) A dynamic evolutionary and functional landscape of plant phased small interfering RNAs. BMC Biol 13(1):1

Chapter 15

Evaluation of Genome-Wide Markers and Orthologous Markers in *Brachypodium distachyon*

Gaurav Sablok, Suresh B. Mudunuri, Korneliya Gudys, Kranthi Chennamsetti, G.P. Saradhi Varma, and Miroslaw Kwasniewski

Abstract

Molecular markers play an important role in identifying the species variation, characterizing the genic diversity, and also linking the identified markers to trait of interest. Genome- and transcriptome-derived molecular markers have been widely used to understand the geographical diversity and have also played a major role in the development of high-density linkage maps. In the present protocol, we present a detailed protocol on bioinformatics approaches towards the whole-genome and transcriptome-assisted simple sequence repeats (SSRs) marker mining in *Brachypodium distachyon* and identification of orthologus SSRs and their validation in *Brachypodium* ecotypes. We also present a protocol for the validation of the identified markers.

Key words *Brachypodium distachyon*, Ecotypes, Orthologous SSRs, Perfect and imperfect SSRs, Repeat density

1 Introduction

Microsatellites or simple sequence repeats (SSRs) [1] are defined as short repetitive stretches of 2–6 bp and are represented throughout the genomes. The primary cause of origin of the microsatellites has been attributed to the slippage mechanism [2] and the mutational robustness of these have been widely studied and addressed across plant genomes. Genome-wide expansion and contraction of these SSRs have not only played an important role in understanding the non-repetitive DNA but also has elucidated the understanding of genome evolution in particular non-repetitive DNA [3]. On the contrary, recent reports suggests that slippage generated microsatellites do not originate from the non-repetitive DNA and its repetition pattern vary among the transcribed regions [4]. In plants, microsatellite markers have played an important role in estimating the species diversity at the interspecies or intraspecies level [5]. In addition to addressing the species diversity, SSRs have

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_15,

played an important role in the characterization of diversity among breeding populations, localization of quantitative trait loci (QTL) and also identifying markers linked to conserved domains. Several efforts have used to identify SSRs from genomes or transcriptomes of plants species and have been used as markers for mapping population [6]. Recently, genic based SSRs, i.e., SSRs located within the coding regions have been used as makers for wide array of applications. Here, we describe a complete protocol for the genome-wide detection of SSRs and the development of the orthologous SSRs across *Brachypodium distachyon* and associated ecotypes.

2 Materials

1. A FASTA file containing the genome sequence. An expert of the FASTA file is given below:

```
>Chromosome1
AGGTTTGTCTTGTGTCTCCGAAAGAAATGGTCCATTTTTTGATGGGCGAACGACGGG-
GATTGAACCCGCGCGTGGTGGATTCACAATCCACTGCCTTGATCCACTTGGCTA-
CATCCGCCCCTACCCCCACACAGGTTTAAGTCTCCATCTACGATCAAGATCATTTCAAA-
TAGAACGAAAATAAAGGAGCAATGGGGTTATTGCTCCTTTATTTTCGTTCTATTTGAAA-
TATATAAATATTCGATTTTCAAAAACTCCTCTACACTAATATACACAAAGAACAAGTCT-
TATCCATTTGTTGGAGCTTCGATAGCAGCTAGATCTAGAGGAAAGTTATGAGCAT-
TACGTTCATGCATAACTTCCATACCAAGGTTAGCACGGTTAATAATATCAGCCCAAGTAT-
TAATTACACGGCCTTGA
```

2. A FASTA containing the transcriptome of ecotypes or related species (*see* **Note 1**).
3. Local version of Imperfect Microsatellite Extractor (IMEx) [7] available from http://www.mcr.org.in/IMEX/download.html, perl version 5.14 (*see* **Note 2**).
4. Download CandiSSRs [8] from http://www.plantkingdomgdb.com/CandiSSR/
5. Download *Brachypodium distachyon* ecotype transcriptome from https://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA182761
6. DNeasy Plant Mini Kit (Qiagen).
7. dNTP Mix (Promega).
8. 10× PCR buffer B (EURx).
9. Color Taq DNA Polymerase (EUR_X).
10. 0.5× TBE buffer.
11. Agarose GP7 Routine (GenoPlast Biochemicals).
12. 100 bp DNA Ladder (Fermentas).
13. Ethidium Bromide Solution (Bio-Rad).

3 Methods

1. **Breakdown of the genome file:** The first step requires the genome breakdown into separate file containing the chromosome level genome sequence. The code to achieve this is enclosed below:

```
#!/usr/bin/perl -w
use warnings ;
($if) = @ARGV ;
if (scalar @ARGV != 1) {
print "ProgName [Input File] \n" ; exit ; }
open(fh1,"$if") ;
@fd1 = <fh1> ;
close(fh1) ;
$i = -1 ; $j = 0 ;
foreach $l (@fd1) {
$i++ ;
if ($l =~ /^>/) {
$j++ ;
$fn = "$j".".inp" ;
write_seq_c($i,$fn) ; } }
exit ;
sub write_seq_c {
($p,$fn1) = @_ ;
open (fh2 , ">$fn1") ;
print fh2 $fd1[$p] ;
for ( $k = $p + 1 ; $k &amp;lt; $p + 100000000 ; $k++) {
if ($k >= (scalar @fd1))
{ last ; }
if ($fd1[$k] eq "")
{last ; }
if ($fd1[$k] =~ />/) { last ; }
print fh2 $fd1[$k] ; }
close(fh2) ;
}
```

2. **Identification of the genome-wide SSRs: Copy the chromosomes files to the IMEx folder and run IMEx on each of the chromosome separately using the following parameters:**

```
$ ./imex_batch chr1.fna 1 1 1 2 2 3 10 10 10 10 10 10 12 6 4 3 3
3 10 1 1 0 10 0 0
```

The above command extracts SRRs of all sizes with a maximum of 10% imperfection that are repeated at least n number of times. The values of *n* are set as follows: Mono—12; Di—6; Tri—4; Tetra-Hexa—3.

SYNTAX:./imex_batch <filename> <mismatches allowed in 6 types> <imperfection % of 6 types> <repeat numbers of 6 types> <flanking sequence length> <align_flag> <text_flag> <coding_flag> <compound_flag> <standardization level> <repsize_flag> [coding_file] (*see* **Note 1**).

3. **For the identification of the orthologous SSRs, define the CTRL file in the CandiSSRs as**:

   #Name Path_to_data_files

   Ref ~/home/ CandiBrachy/Bdistachyon_314_v3.1.fa

   Gre ~/home/ CandiBrachy/Brasy-Gre_cDNA.fa

   Esp ~/home/ CandiBrachy/Brasy-Esp_cDNA.fa

   Cor ~/home/ CandiBrachy/Brasy-Cor_cDNA.fa

   CandiSSRs can then be run as $ perl CandiSSR.pl -i Ctg_file -o CandiSSR_Run -p Prefix -l FlankingLen -s Identity -c Coverage -t Cpu (*see* **Note 2**).

   -o <str> Name of directory for output. [default: CandiSSR_Run]
   -p <str> The prefix of output file. [default: CandiSSR_Output]
   -l <int> The flanking sequence length of SSRs. [default: 100]
   -e <int> Blast evalue cutoff. [default: 1e-10]
   -s <int> Blast identity cutoff. [default: 95]
   -c <int> Blast coverage cutoff. [idefault: 95]
   -t <int> Number of CPU used in blast searches. [default: 10]

4. DNA isolation**:** Approximately 100 mg of tissue from fresh young leaves of each genotype was homogenized in a sterile mortar. Genomic DNA was isolated using a DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions. The yield and purity of the DNA was determined using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific), and the samples were diluted in sterile water to 50 ngμl^{-1}.

5. PCR amplification and detection of SSR bands.

   SSRs amplification: SSR amplification reactions were performed using thermocycler (Biometra) in a 20 μl reaction volume containing 5 ng of genomic DNA, 1× PCR buffer, 0.25 mM dNTPs (Promega), 2.5 μM of each primer (Sigma-Aldrich), and 0.05 U Color Taq polymerase (EURx). PCR were held at 94 °C for 3 min, followed by 35 cycles of 94 °C for 30 s, 56 °C for 20 s and 72 °C for 20 s, with a final extension step at 72 °C for 5 min. The PCR products were separated on 4% agarose gel (GenoPlast Biochemicals) stained with ethidium bromide (Bio-Rad) in 0.5× TBE buffer at constant voltage

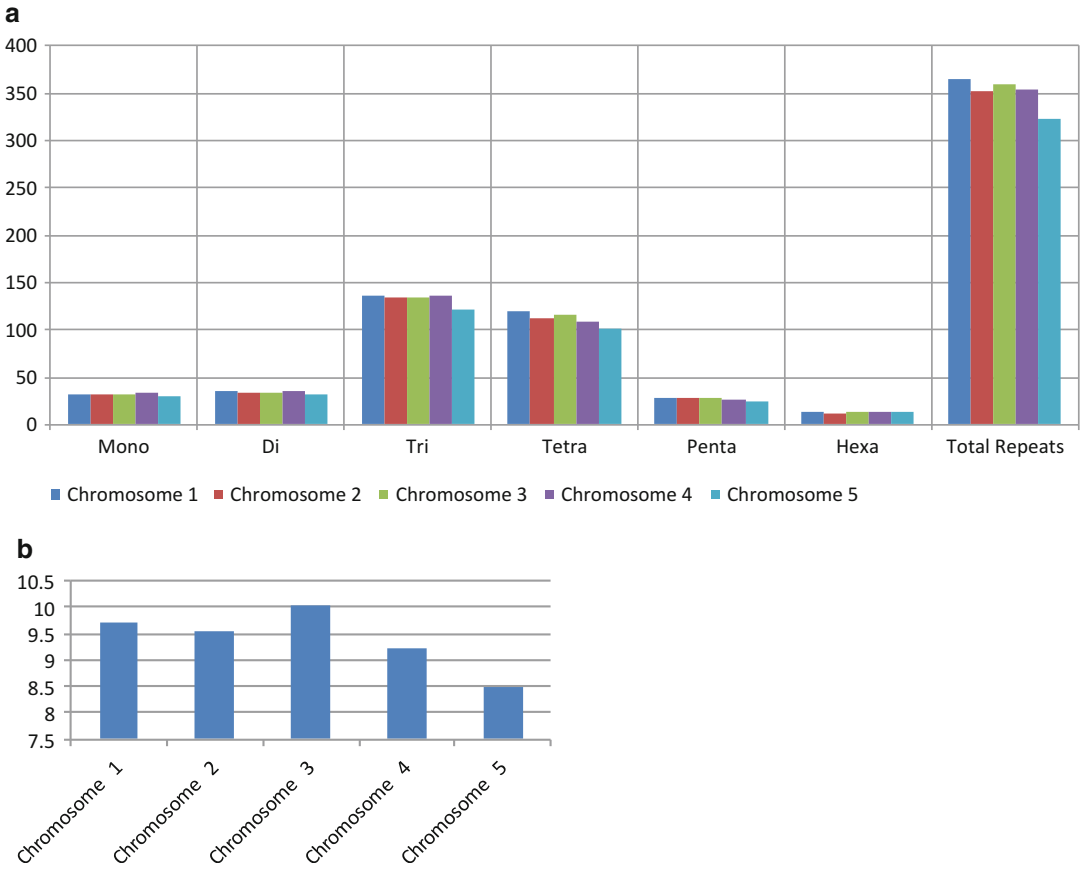

Fig. 1 (**a**) Perfect repeats density; (**b**) Compound perfects repeats

5 V/cm. Number of alleles per locus and allele sizes were determined in GelAnalyzer software using 100 bp DNA ladder (Fermentas) as a reference point.

6. Figure 1a, b shows the perfect repeat density and the distribution of compound perfect repeats across the *Brachypodium distachyon* genome which is defined as the number of microsatellites per MB of genome. Figure 2a, b represents the imperfect repeat density and the distribution of compound imperfect repeats across the *Brachypodium distachyon* genome. Perfect repeats represent the continuous stretches of the defined nucleotides whereas the imperfect repeats represent two SSRs represented by an imperfect unit of SSRs. The minimum imperfection defined as 10.
7. Figure 3 represent the validation of the orthologous SSRs across *Brachypodium distachyon* ecotypes.

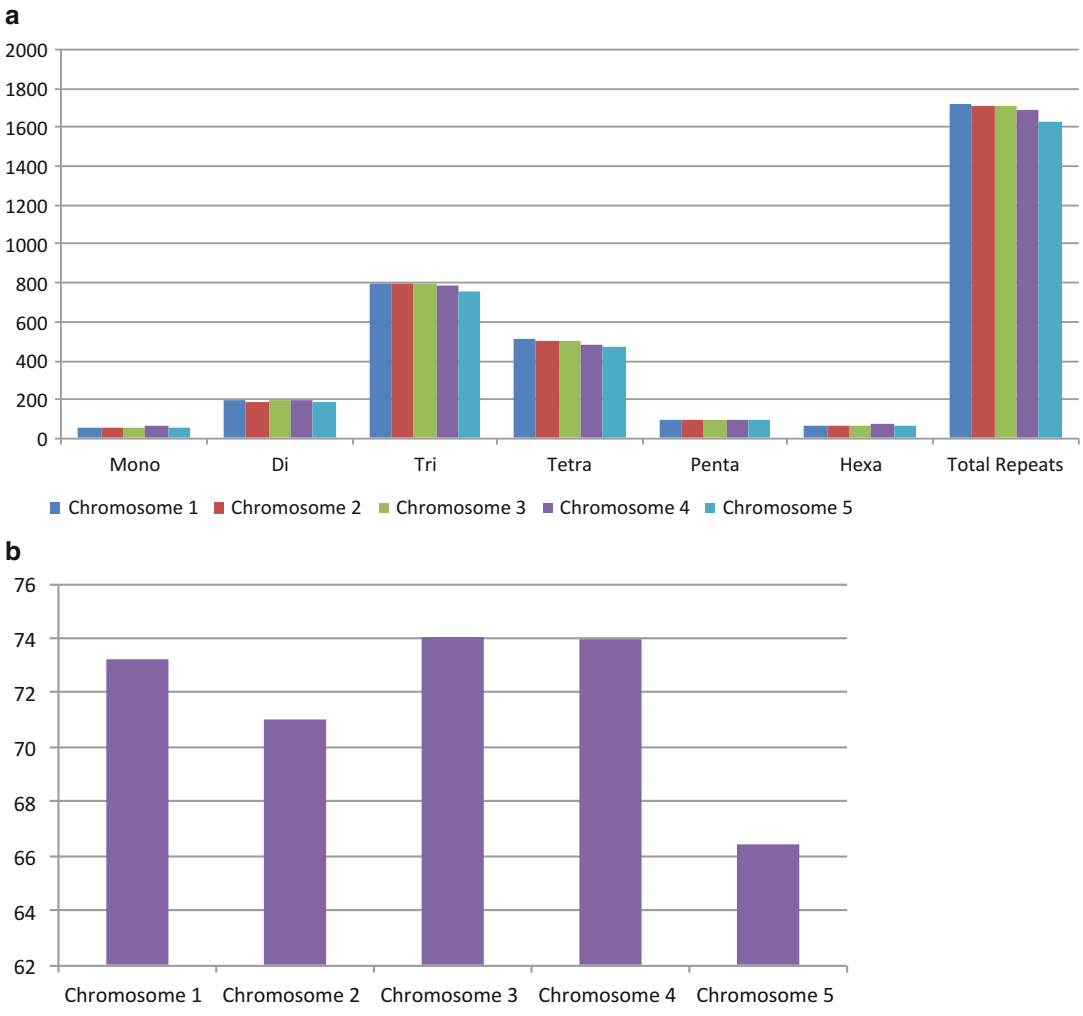

Fig. 2 (**a**) Imperfect repeats density; (**b**) Compound imperfect repeats

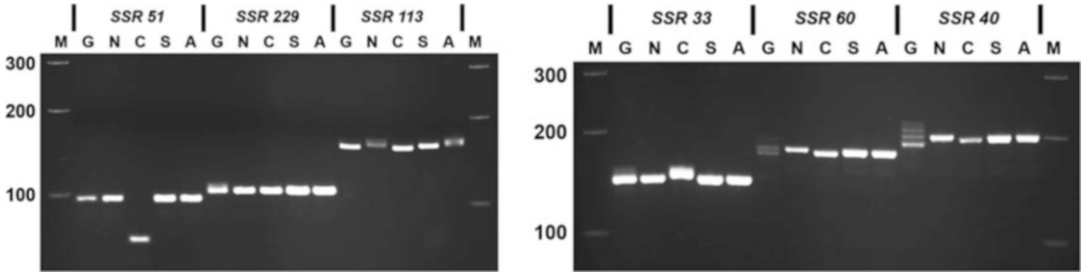

Fig. 3 Validation of the orthologous SSRs

4 Notes

1. In case of genome-wide detection of SSRs, genome should be available as FASTA file with no spaces in the header. Additionally, the coding sequences and the protein information can be supplied in the form of the PTT table to link the identified SSRs to the coding regions.
2. In case of transcriptomes, the assembly should represent a nonredundant assembly with the availability of similar species or the species sharing the same clade with related closeness.

References

1. Ellegren H (2014) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445
2. Leclercq S, Rivals E, Jarne P (2010) DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. Genome Biol Evol 2:325–335
3. Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200
4. Weitzman JB (2002) Microsatellites in plant genomes. Genome Biol 3:spotlight-20020128-01
5. Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN et al (2011) Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in Brachypodium. PLoS One 6(6):e21298
6. Xiao Y, Xia W, Ma J, Mason AS, Fan H, Shi P, Lei X, Ma Z, Peng M (2016) Genome-wide identification and transferability of microsatellite markers between Palmae species. Front Plant Sci 7:1578
7. Suresh BM, Nagarajaram HA (2007) IMEx: imperfect microsatellite extractor. Bioinformatics 23:1181–1187
8. Xia E-H, Yao Q-Y, Zhang H-B, Jiang J-J, Zhang L-P, Gao L-Z (2016) CandiSSR: an efficient pipeline used for identifying candidate polymorphic SSRs based on multiple assembled sequences. Front Plant Sci 6:1171

Chapter 16

Protocol for Coexpression Network Construction and Stress-Responsive Expression Analysis in *Brachypodium*

Sanchari Sircar, Nita Parekh, and Gaurav Sablok

Abstract

Identifying functionally coexpressed genes and modules has increasingly become important to understand the transcriptional flux and to understand large scale gene association. Application of the graph theory and combination of tools has allowed to understand the genic interaction and to understand the role of hub and non-hub proteins in plant development and its ability to cope with stress. Association genetics has also been coupled with network modules to map these key genes as e-QTLs. High throughput sequencing approaches has revolutionized the mining of the gene behavior and also the association of the genes over time-series. The present protocol chapter presents a unified workflow to understand the transcriptional modules in *Brachypodium distachyon* using weighted coexpressed gene network analysis approach.

Key words *Brachypodium distachyon*, Drought stress, Functional modules, Network analysis, Co-expression analysis

1 Introduction

Transcriptional gene regulation plays an important role in understanding the genic behavior of plant species to cope with the corresponding stress response. Being, sessile, plants respond to the physiological stress through the coordinated regulation of genes and interactions among the genes. With the advent of the next generation sequencing, plethora of experimental models have been widely studied to identify stress regulators using different combination of approaches such as RNA-seq, microarray, ChIP-Seq and Methyl-Seq. By incorporating these approaches, several questions toward understanding the system biology of stress have been widely addressed. Most importantly, application of these

Electronic supplementary material: The online version of this chapter (doi:10.1007/978-1-4939-7278-4_16) contains supplementary material, which is available to authorized users.

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_16, © Springer Science+Business Media LLC 2018

approaches with high throughput computational modeling approaches has widely revealed the interactions among stress regulators. Among these system biology approaches, application of network modeling is gaining importance. High throughput application of network biology has revealed the role of graph theory to understand large scale gene association. This approach have been used on a range of systems such as biotic and abiotic stress responses using whole transcriptome sequencing in potato [1], understand seed germination in *Arabidopsis* [2], identification of novel drought-responsive genes in rice [3], transcriptional reprogramming in *Arabidopsis* due to mechanical wounding and insect herbivores [4], etc. In this protocol chapter, we describe a step-by-step protocol for usage of network biology to understand the modules in *Brachypodium distachyon*.

1.1 WGCNA

In this chapter, we describe a step-by-step protocol for finding clusters (modules) of highly correlated genes using weighted gene correlation network analysis (WGCNA), an R package [5] in response to drought stress in *Brachypodium distachyon*. In general, a coexpression network is constructed by considering genes as *nodes* and connections between them represented by *edges* based on some measure of association. Mathematically, a network is represented by an adjacency matrix, $A = [a_{ij}]$, where the elements a_{ij} represent the strength of association between gene pairs i, j in a weighted network, while it takes values "1" or "0" for an unweighted network. In WGCNA, the measure of association, i.e., the coexpression similarity s_{ij} is defined as the correlation between the expression profiles (x) of genes i and j:

$$s_{ij} = |\ \mathrm{cor}(x_i, x_j)\ | \tag{1}$$

This correlation matrix s_{ij} is converted into an adjacency matrix a_{ij}. For unweighted network, adjacency between gene expression profiles of x_i and x_j is defined by hard-thresholding the coexpression similarity s_{ij} as:

$$a_{ij} = \begin{cases} 1 \ if \ s_{ij} \geq \tau; \\ 0 \ otherwise, \end{cases} \tag{2}$$

where τ is the "hard" threshold parameter. While unweighted networks are widely used, they do not reflect the continuous nature of the underlying coexpression information and thus lead to an information loss. In contrast, weighted networks allow the adjacency to take continuous range of values between 0 and 1. By raising the coexpression similarity to a power:

$$a_{ij} = s_{ij}^{\beta} = \left|\mathrm{cor}(x_i, x_j)\right|^{\beta} \tag{3}$$

where the soft-threshold parameter $\beta \geq 1$. The weighted gene coexpression network construction emphasizes high correlations

at the expense of low correlations. In the weighted network, every node is connected to every other node with varying correlation strengths given by a_{ij}. The module detection then identifies clusters of densely interconnected genes. Analysis of coexpressed modules helps us to focus on biological processes/pathways represented by the clusters of genes instead of individual genes. Most standard clustering methods require a distance, or dissimilarity measure, to obtain clusters. Highly coexpressed genes have a small dissimilarity which, in WGCNA, is defined by the topological overlap matrix (TOM) as

$$\text{dissTOM}_{ij} = 1 - \text{TOM}_{ij} = 1 - \frac{\sum_{u \neq i} a_{iu} a_{uj} + a_{ij}}{\min\left(k_i, k_j\right) + 1 - a_{ij}} \quad (4)$$

where $k_i = \sum_{u \neq i} a_{ui}$ denotes the network connectivity. TOM combines the connection strength between a pair of genes with their connections to other "third party" genes, and has been shown to be a highly robust measure of network interconnectedness. The dissimilarity measure is then used as an input for average linkage hierarchical clustering, wherein, the closest nodes are merged iteratively to obtain a hierarchical clustering tree (dendrogram) that provides information on how the genes are merged together. Modules are then defined as branches of the resulting tree.

2 Materials

For the construction and analysis of correlation network, we consider gene expression data from the leaf tissue of *Brachypodium distachyon* under drought stress by Verelst et al. [6]. In their study the authors performed Affymetrix tiling array analysis of the transcriptomes from different developmental zones on the leaf tissue, viz., proliferating zone (P), expanding zone (E), and mature zone (M). Three biological replicates from each of the developmental zones were sampled in control and severe drought stress and two biological replicates were sampled in mild drought stress condition. In this protocol, we consider the samples corresponding to the control and severe drought stress condition.

For the illustration of network construction and analysis, the SOFT file for *Brachypodium distachyon* under drought stress by author et al. (GSE38247) is downloaded from the NCBI-GEO. This file has RMA-normalized expression values for 18 samples, three controls and three corresponding to severe drought condition for each of the three stages, viz., proliferation, expansion and maturation. The dataset consists of expression values of 27,699 probes in the SOFT file which is now copied to a spreadsheet with the probe ids as rows and the 18 samples as columns to create

a tab-limited text file, "drought_rma.txt" in the working directory. A phenotypic data file "pData.txt" containing the sample information (case/control status) is prepared as shown below. Care should be taken to see that the sample names are defined in the same manner as in the "drought_rma.txt" file. For example, the sample names in this file are represented as P.control1 for the first replicate from the proliferation zone corresponding to the control state and P.severe1 corresponding to the first replicate in the severe drought state. Similarly, the samples are labeled with the prefix E for the expansion and M for the mature stages. Thus, the phenotypic data file, "pData.txt" is prepared in the following tab-separated format:

```
> head(pData)
           type       status
P.control1 P_control control
P.control2 P_control control
P.control3 P_control control
P.severe1     P_case    case
P.severe2     P_case    case
P.severe3     P_case    case
⋮
```

Here, the first column refers to the sample names, the second and third column refer to any trait information, such as "type" and "status".

2.1 R Packages for Microarray Analysis

In this protocol we will show network construction and analysis using freely available software packages in R. The first step is to identify and install various R packages required for the network analysis. Here we will start with installing two BioConductor packages in R, "simpleaffy" and WGCNA R. Before installing these packages, the user needs to initiate the R environment from the working directory (by typing the command R). The "simpleaffy" package provides functionalities to read Affymetrix .CEL files, preprocessed expression files, and phenotypic data; and computes simple measures such as fold change and p-values for t statistics [7]. It can be installed by the following R command:

```
# install Bioconductor package
>source("https://bioconductor.org/biocLite.R")
> biocLite("simpleaffy")
```

The WGCNA R package consists of various R functions $s_{ij} = |\mathrm{cor}(x_i x_j)|$ for construction and topological analysis of weighted coexpression networks. The information about the latest version and various dependencies can be obtained from the link [ ].

WGCNA can be installed from Comprehensive R Archive Network (CRAN) using the following command:

```
>biocLite(c("AnnotationDbi", "impute", "GO.db", "preprocessCore"))
>install.packages("WGCNA")
```

3 Methods and Results

3.1 Data Preprocessing

The first step before doing any analysis is to see if there are no outliers in the data, and remove them, if any. Boxplots are frequently used to obtain a distribution of the data along with the central values and variability (quartile ranges) across the samples to detect possible outliers. The user can obtain Boxplots for each sample in the file "drought_rma.txt" using the following commands in R:

```
>raw_data<-read.table('drought_rma.txt', header=TRUE, sep='\t')
> boxplot(raw_data[,2:19])# the data starts from column 2 and ends in column 19
```

In Fig. 1 is shown the Boxplots showing the distribution of normalized, $\log_2$-transformed expression values for the 18 samples. It is clear that there are no probable outliers and the distribution of the data across the samples is more or less the same.

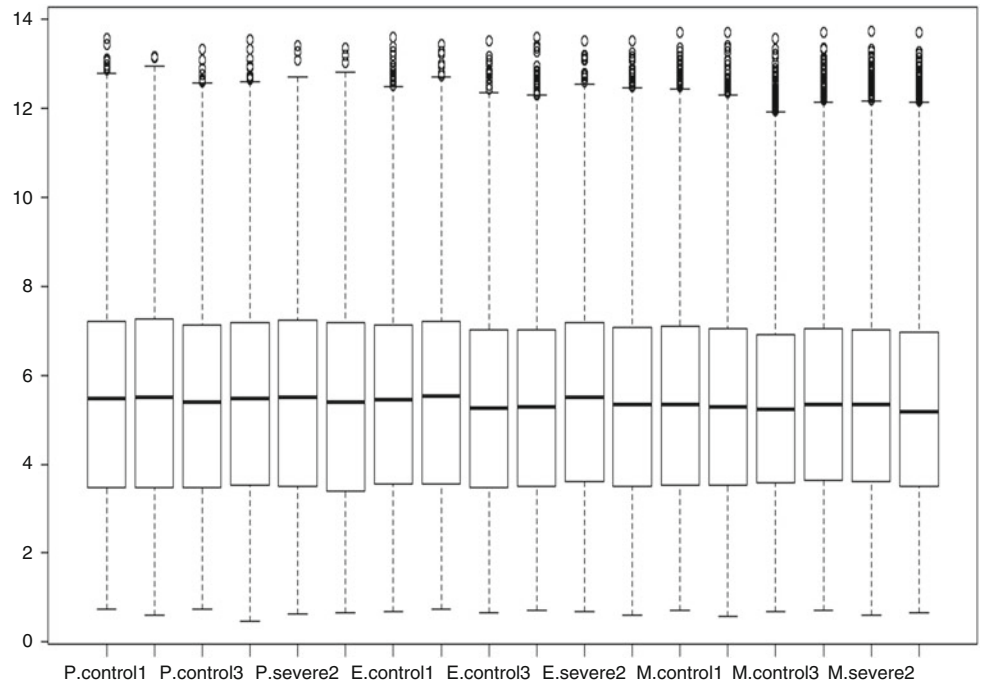

Fig. 1 Boxplot showing the distribution of normalized, $\log_2$-transformed expression values for the 18 samples

Next, the 27,699 probes in the GSE38247 dataset are mapped to the ensemble gene ids fetched using BioMart in EnsemblPlants [8]. Since, in general, multiple probes map to the same gene, the expression for each gene is considered by either taking an average of the expression values of the probes across all the samples, the probe with highest coefficient of variation, or with highest fold-change across samples. Here we consider an average of the expression values of the probes mapping to the same gene. This resulted in 24,686 unique probe–gene pairs. Next, genes with very low intensity values (<30 in 70% of the samples) are removed as these result in false indication of differential fold-change. The remaining 12,359 genes are then used for the construction of the correlation network.

3.2 Network Construction Using WGCNA

3.2.1 Data Loading

The expression matrix consisting expression values of 12,359 genes and 18 samples is saved in a file "brachyData_input.csv" in a .csv format. The following commands in R load the expression matrix as input:

```
>library(WGCNA)
>allowWGCNAThreads()
>options(stringsAsFactors = FALSE);
>brachyData = read.csv("brachyData_input.csv");
>head(brachyData)    #how brachyData_input.csv looks like
genes  P.control1  P.control2  P.control3  P.severe1  P.severe2
P.severe3
1 Bradi1g68840       5.3       5.2       5.2       5.2       5.7
5.3
2 Bradi2g07260       4.7       5.0       4.7       4.6       5.0
4.9
3 Bradi1g58330       5.6       5.8       5.5       5.5       5.6
5.8
4 Bradi2g52670       5.8       5.7       5.2       5.5       5.6
4.6
5 Bradi3g59630       8.5       8.4       8.1       8.2       8.3
8.0
6 Bradi2g42160       6.7       6.9       6.5       6.5       6.8
6.6
   E.control1  E.control2  E.control3  E.severe1  E.severe2
E.severe3 M.control1
1        6.3        6.2        6.1       6.2       6.8       6.3
7.6
2        5.6        5.7        5.5       5.1       5.8       5.4
6.6
3        4.9        5.0        5.0       5.2       5.0       5.0
4.9
4        5.9        6.2        6.0       6.0       6.2       5.7
6.7
5        7.3        7.5        7.1       7.3       7.4       7.1
```

```
7.6
6        6.3        6.7        6.2        6.4        6.4        6.3
6.4
  M.control2 M.control3 M.severe1 M.severe2 M.severe3
1        7.7        7.2        7.1        7.5        7.5
2        6.4        6.2        6.1        6.5        5.9
3        5.0        4.8        5.3        5.0        4.9
4        6.5        5.7        6.8        7.1        6.7
5        7.4        7.0        7.5        7.5        7.3
6        6.6        6.0        6.3        6.5        6.2
>datExpr0 =as.data.frame(t(brachyData[, -c(1)]))
>names(datExpr0) = brachyData$genes;
>rownames(datExpr0) =names(brachyData)[-c(1)];
# A check for excessive missing values
>gsg = goodSamplesGenes(datExpr0, verbose = 3);
>gsg$allOK
```

If the last statement returns TRUE, then the file has been read correctly and all genes have passed the validation process.

3.2.2 Selection of the Soft-Thresholding Power, β

The coexpression similarity matrix is scaled using soft-thresholding power β based on the criterion of approximate scale-free topology. After the data has been read properly, the function "pickSoftThreshold" is used to choose a proper soft-thresholding power for network construction. For the analysis of network topology, two graphs, one corresponding to the scale-free topology model fit, and the other mean connectivity (degree), are both plotted as a function of soft-thresholding (power) and must be carefully inspected by the user to choose a proper value of soft-threshold power, β. The following R commands generate the graphs (given in Fig. 1) that aid in choosing a proper soft-thresholding power. The user chooses a set of candidate powers (the function provides suitable default values), and the function returns a set of network indices that should be inspected, as shown below:

```
>powers =c(c(1:10),seq(from = 12, to=40,by=2))
# Function to compute soft-threshold
>sft = pickSoftThreshold(datExpr0, powerVector = powers,
verbose = 5)
# plotting of the results
>sizeGrWindow(9, 5)
>par(mfrow =c(1,2));
>cex1 = 0.75
>plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fi-
tIndices[,2],
      xlab="SoftThreshold(power)",ylab="Scale Free Topology
Model Fit, >signed R^2", type="n",
    main =paste("Scale independence"));
```

```
>text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fi-
tIndices[,2],>labels=powers,cex=cex1,col="red");>abline
(h=0.75,col="red")# draw a line at R-squared value 0.75
>plot(sft$fitIndices[,1], sft$fitIndices[,5],
      xlab="SoftThreshold(power)",ylab="Mean Connectivity",
type="n",
      main =paste("Mean connectivity"))>text(sft$fitIndices
[,1], sft$fitIndices[,5],labels=powers, cex=cex1,col="red")
```

The coefficient of determination, R^2 value, is used to evaluate the scale-free topology model fit, and it ranges from 0 (indicating no linear relationship between the variables) to 1 (indicating a perfectly linear relationship). Based on the distribution of the R^2 value across the soft-thresholding powers, we choose a cutoff for the scale-free fit model as 0.75 for this dataset.

From Fig. 2 it is seen that the mean connectivity is too high at $\beta = 6$ or 7. Therefore, a higher power, $\beta = 22$, is selected. The details of the mean, median, and maximum connectivity (k) for $\beta = 22$ are given in Table 1.

3.2.3 *Module Detection*

Once the parameters for network construction have been chosen, the next step is the detection of modules of coexpressed genes. WGCNA provides three different options for module detection: a single-step network and module detection construction (for first

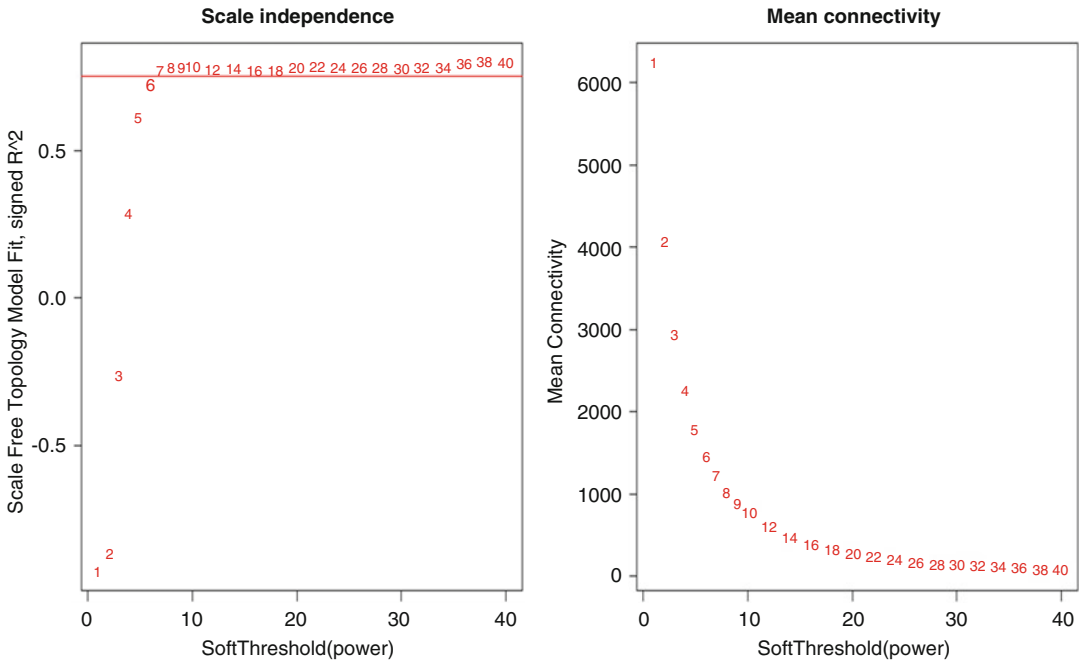

Fig. 2 Analysis of network topology for various soft-thresholding powers. In left *panel* is shown the scale-free fit index, and *right panel*, the mean connectivity (degree) as a function of the soft-thresholding power

Table 1
Network Parameters for β = ss22

Soft-threshold	R^2 value	Mean *k*	Median *k*	Max *k*
22	0.78	217	81.7	1110

time users), a detailed step-by-step network construction and module detection method (for users to experiment with alternate options) and a block-wise network construction and module detection method (to analyze large datasets). Here we use the single-step network construction method using the following R commands:

```
>net = blockwiseModules(datExpr0,
maxBlockSize = 13000,
power= 22, #selected from scale-free fit
TOMType ="unsigned", # unsigned network is constructed
deepSplit = 3, # default is 2
minModuleSize = min(20, ncol(datExpr)/2 )
# p-value ratio threshold for reassigning genes between
modules
reassignThreshold = 0,
mergeCutHeight = 0.25,
#modules be labeled by colors (FALSE), or by numbers (TRUE)
numericLabels = FALSE,
pamRespectsDendro = FALSE,
saveTOMs = TRUE, #save Topological Overlap Matrix
saveTOMFileBase ="brachyData_TOM",
>verbose = 3)
```

The "blockwiseModules" function has a number of parameters which can affect the network construction, such as number and size of the modules detected. The user needs to set the parameter "maxBlockSize" depending on the number of probes/genes in the dataset. Since in our dataset we have 12,359 number of genes, here it has been set to 13,000 (default 5000). The user needs to take into consideration the memory of the system when the dataset size >5000 probes. The user is advised to either use a high-end system (32 GB) or use the block-wise network construction and module detection method. The parameter "deepSplit" controls the number and size of the modules detected. It ranges from 0 to 4 and for smaller value we get fewer large modules while for higher values, the number of modules increase and the size decreases. Here, we have considered deepSplit = 3 (default 2). The parameter "mergeCutHeight" (dendrogram cut-height) is the threshold for merging the modules and has been set to 0.25 (default 0.15). The TOM matrix is saved as "brachyData_TOM".

To see the modules that have been created and their size, one can use the command:

```
# number of modules and their sizes:

> table(net$colors)
black      blue     brown     green
153        3275      1451     278

grey    magenta    pink    purple
1072    62         143     36

red    turquoise    yellow
220    4734        935
```

We observe that there are ten modules labeled by color, ordered with decreasing size ranging from 4734 to 36 genes. The module "grey" is reserved for genes that do not belong to any of the coexpressed modules.

The hierarchical clustering dendrogram used for module detection can be displayed with color assignment using the following R commands. The resulting plot is shown in Fig. 3.

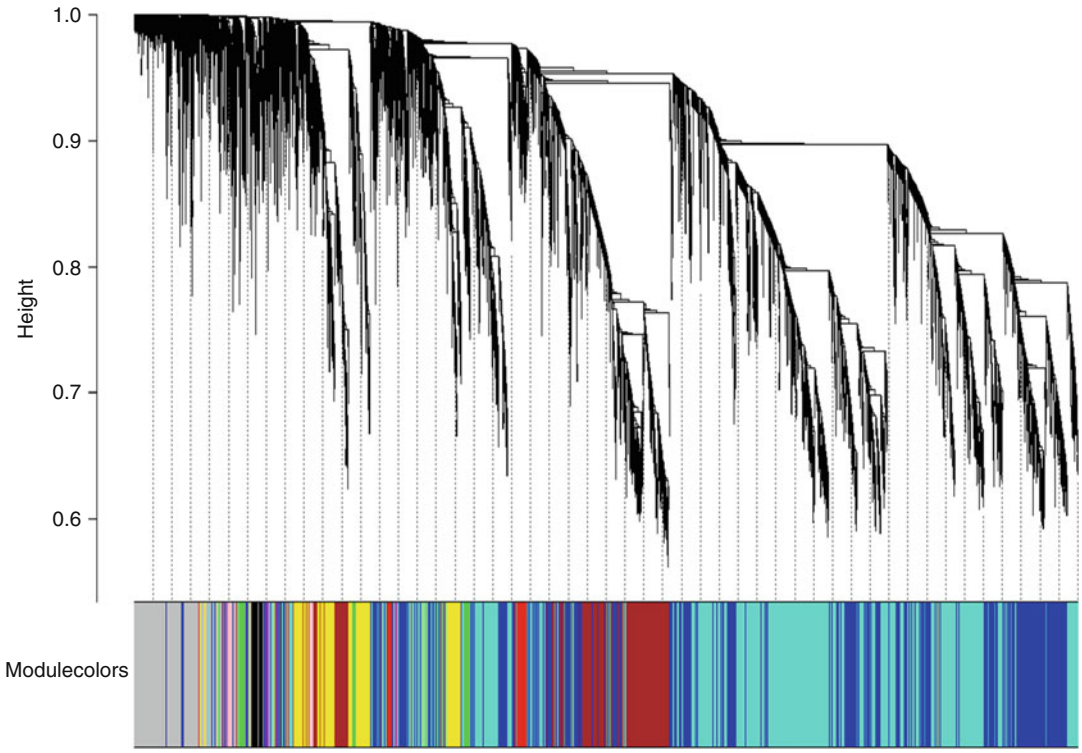

Fig. 3 Clustering dendrogram of genes, with dissimilarity based on topological overlap, together with assigned module colors

```
# plotting the dendogram:
>sizeGrWindow(12, 9)
>mergedColors = labels2colors(net$colors)
>plotDendroAndColors(net$dendrograms[[1]], mergedColors[net
$blockGenes[[1]]],
"Modulecolors",
dendroLabels = FALSE, hang = 0.03,
addGuide = TRUE, guideHang = 0.05)
#saving the modules with the color assignments:
>moduleLabels = net$colors
>moduleColors = labels2colors(net$colors)
>MEs = net$MEs;
>geneTree = net$dendrograms[[1]];
>save(MEs, moduleLabels, moduleColors, geneTree,fi-
le="drought_gene_tree_v1.RData")
```

3.2.4 Centrality Measures

To identify the importance of a gene within the module/whole network, various measures such as intramodular connectivity, module membership, and eigengene-based connectivity can be computed. The intramodular connectivity (kIN) is a measure of the connectivity of a gene within a module. For weighted networks, it is the sum of the adjacencies of a gene within the module. The following R commands may be used to compute kIN:

```
#extract intramodular connectivity, extra-modular connectiv-
ity, and the difference of the intra- and extra-modular
connectivities
> ADJ1=abs(cor(datExpr0,use="p"))^22
>Alldegrees1=intramodularConnectivity(ADJ1, moduleLabels)
```

To compute how well a gene is connected to other (biologically interesting) modules, the module eigengene-based connectivity (kME) measure is calculated for each gene i and is defined as the correlation between its expression and the module eigengene. Mathematically,

$$\mathrm{kME}_{\mathrm{turquoise}}(i) = \mathrm{cor}\left(x_i, \mathrm{ME}_{\mathrm{turquoise}}\right)$$

where x_i is the gene expression profile of the gene i and $\mathrm{ME}_{\mathrm{turquoise}}$ is the module eigengene of the brown module. Here, the gene i may not be a part of the turquoise module.

```
#eigengene-based connectivity
>datME=moduleEigengenes(datExpr0,moduleColors)$eigengenes
#default measure of correlation: Pearson
>signedKME(
  datExpr0, datME,
#Extract the module membership  outputColumnName = "MM",
  corFnc = "cor", corOptions = "use = 'p'")
```

3.2.5 Exporting Modules

For analysis and visualization of the coexpressed subnetworks in software programs such as Cytoscape [9], one requires to export the list of genes and their connectivity (edge-list) in each module. For this, a node-list file can be created in a spreadsheet with the node names in the first column and any other gene id/name in the second column and save it as "annotation.csv". For this network, since we have already mapped probes to the gene ids, our node list column and gene symbol column is the same.

```
>annot =read.csv(file="annotation.csv");
>head(annot)genes  gene_symbol
1 Bradi1g68840 Bradi1g68840
2 Bradi2g07260 Bradi2g07260
3 Bradi1g58330 Bradi1g58330
4 Bradi2g52670 Bradi2g52670
5 Bradi3g59630 Bradi3g59630
6 Bradi2g42160 Bradi2g42160
#extract node-list and edge-list one by one using the module names
>modules =c("yellow");
>inModule =is.finite(match(moduleColors, modules));
>modProbes = genes[inModule];
#match node names
>modGenes = annot$gene_symbol[match(modProbes, annot$genes)];>
modTOM = TOM[inModule, inModule];
>dimnames(modTOM) =list(modProbes, modProbes)
#export to Cytoscape
>cyt = exportNetworkToCytoscape(modTOM,
edgeFile =paste("CytoscapeInput-edges-1",paste(modules, col-
lapse="-"),".txt", sep=""),
nodeFile =paste("CytoscapeInput-nodes-1",paste(modules, col-
lapse="-"),".txt", sep=""),
weighted= TRUE,
threshold = 0.02,
nodeNames = modProbes,
altNodeNames = modGenes,
nodeAttr = moduleColors[inModule]);
```

3.2.6 Functional Characterization of the Modules

The most important part of network analysis is the functional characterization of the coexpressed genes and modules. Of particular interest is identifying the biological processes associated with the genes since these describe the physiological roles carried out by the genes. There are two aspects to functional characterization. One involves the functional annotation of individual genes to identify the role of important candidate genes and the other involves functional characterization at the module-level, enrichment analysis of coexpressed modules to identify the significant shared GO terms in the gene set. In WGCNA, enrichment analysis can be performed using the function "GOenrichmentAnalysis" which requires organism-specific annotation packages to be installed using

AnnotationDBI (BioConductor package). However, for Brachypodium, annotation package is not available. In such a situation we need to explore various plant-based resources available online, such as Phytozome [10], Agrigo [11], and Gramene [12]. For the analysis, we retrieve gene descriptions for 12,359 Brachypodium genes using BioMart in Phytozome. Since this is a published dataset, annotations from the authors is also available in the Supplementary Material (Table 3) [6] and is considered by us. For the enrichment analysis of the modules, AgriGo and enrichment analysis tool from the GO Consortium were used.

3.3 Identification of Differentially Expressed Genes

To identify the drought-responsive modules, we map the differentially expressed genes (DEGs) to the ten modules obtained above. For identifying the differentially expressed genes, the R package "simpleaffy" is used. The same input file, brachyData_input.csv that was used for WGCNA, is used as input along with the phenotypic data file, "pData.txt".

```
>library(simpleaffy)
#load the expression matrix
>exprs <-as.matrix(read.table(brachyData_input.csv', header=TRUE, sep=',', row.names=1, as.is=TRUE))
#load the phenotypic data
>pData<-read.table('pData.txt', row.names=1, header=TRUE, sep="\t")
>metadata<-data.frame(labelDescription=c("zones","case/cntrol"), row.names=c("type","status"))
>phenoData<-new("AnnotatedDataFrame", data=pData, varMetadata=metadata)
#create an expressionset object
>exampleSet<- ExpressionSet(assayData=exprs, phenoData=phenoData)
# fold change and t test for proliferation stage
>p_zone<-get.fold.change.and.t.test(exampleSet,"type",c("P_case","P_control"))
>write.table(fc(p_zone), "p_fc.txt")
>write.table(tt(p_zone), "p_tt.txt")
# fold change and t test for expansion stage
>e_zone<-get.fold.change.and.t.test(exampleSet,"type",c("E_case","E_control"))
>write.table(fc(e_zone), "e_fc.txt")
>write.table(tt(e_zone), "e_tt.txt")
# fold change and t test for maturation stage
>m_zone<-get.fold.change.and.t.test(exampleSet,"type",c("M_case","M_control"))
>write.table(fc(m_zone), "m_fc.txt")
>write.table(tt(m_zone), "m_tt.txt")
```

Table 2
Differentially expressed genes across the three stages

	DEGs in the three stages			Across three stages	
Total sof DEGs	Proliferation	Expansion	Maturation	Upregulated	Downregulated
2471	952 U: 462, D: 490	988 U: 396, D: 591	1136 U: 494, D: 642	46	52

Here, the data objects "p_zone", "e_zone", and "m_zone" contain the fold change and *t* test results and are written in the respective files "p_fc.txt", "p_tt.txt", and so on. These text files are opened using a spreadsheet and genes are filtered as differentially expressed if the *p*-value <0.05 and fold change >user-defined threshold (1.2 in this case). The filtered set of differentially expressed genes (DEGs) is then mapped to the modules to identify modules enriched with the DEGs. In Table 2 is shown the number of upregulated (U) and downregulated (D) genes across the three stages. We observe that the number of downregulated genes is more than that of upregulated genes across all the three stages. The number of DEGs shared among the three stages is very low suggesting different processes are initiated at different developmental stages in leaf growth in response to drought.

In Table 3 is shown the various modules along with the enriched GO terms. The percentage of DEGs in each developmental stage is indicated with the absolute number of up and downregulated genes.

From Table 3 it is observed that the majority of DEGs are mapped to the Purple, Black, and Magenta modules which are associated with regulation of cellular biosynthetic processes, nitrogen compound metabolic processes, and tRNA metabolic processes. For the proliferation stage, the largest percentage of DEGs is mapped to the magenta module. This cluster may represent a novel functional module as ~50% of the genes are unclassified.

Further downstream analysis of the drought-responsive modules may help in identifying key genes such as transcription factors which may be induced due to drought stress. Moreover, coexpressed genes which are densely connected to each other in the module are likely to share similar *cis*-elements. Analysis of such clusters can help in identifying important signaling pathways. Topological analysis of the modules using various centrality measures like intramodular connectivity and eigengene-based connectivity will help in screening important candidate genes in the modules.

Table 3
Coexpressed modules with enriched go terms and %age of DEGs across the three stages

	%age DEGs			GO enrichment results	
Module	**Proliferation**	**Expansion**	**Maturation**	**AgriGo**	**GO consortium**
Black (153)	12 (U: 5, D: 14)	28.7 (U: 5, D: 39)	27.4 (U: 7, D: 35)	ncRNA metabolic process, tRNA metabolic process, RNA metabolic process	Plastid organization, ncRNA metabolic process, chloroplast organization, tRNA metabolic process
Blue (3275)	7 (U: 120, D: 110)	6.8 (U: 98, D: 126)	9.1 (U: 94, D: 204)	Metabolic process, nitrogen compound metabolic process, photosynthesis	Glyceraldehyde-3-phosphate metabolic process, cellular aldehyde metabolic process, small molecule metabolic process, isoprenoid metabolic process
Brown (1451)	5.7 (U: 56, D: 26)	8.4 (U: 26, D: 96)	9.6 (U: 79, D: 60)	Cellular process, vesicle-mediated transport, establishment of localization, establishment of localization	Vesicle-mediated transport, gene expression, intracellular transport, cellular localization
Green (278)	2 (U: 3, D: 3)	2.5 (U: 5, D: 2)	2.5 (U: 4, D: 3)	Chromosome organization, cellular component assembly, chromatin organization, cellular component organization	Single-organism metabolic process, single-organism process, single-organism cellular process, organic substance metabolic process
Magenta (62)	79 (U: 43, D: 6)	56 (U: 30, D: 5)	46.8 (U: 24, D: 5)	–	Unclassified
Pink (143)	2.8 (U: 4, D: 0)	6.3 (U: 8, D: 1)	2.1 (U: 2, D: 1)	No enrichment for biological process; molecular function (binding, catalytic activity, ATPase activity)	Unclassified
Purple (36)	69 (U: 17, D: 8)	72.2 (U: 18, D: 8)	52.8 (U: 15, D: 4)	Regulation of cellular biosynthetic process, nitrogen compound metabolic process, regulation of	Unclassified

(continued)

Table 3
(continued)

Module	%age DEGs			GO enrichment results	
	Proliferation	Expansion	Maturation	AgriGo	GO consortium
				transcription, regulation of gene expression	
Red (220)	0	0.45 (U: 1, D: 0)	0	Regulation of transcription, regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	Single-organism process, single-organism cellular process, single-organism metabolic process, cellular process
Turquoise (4734)	8.3 (U: 121, D: 271)	7.7 (U: 145, D: 219)	8.8 (U: 191, D: 224)	Metabolic process, cellular metabolic process, primary metabolic process, gene expression, translation	Gene expression, cellular nitrogen compound metabolic process, nitrogen compound metabolic process, organonitrogen compound biosynthetic process
Yellow (935)	9.3 (U: 50, D: 37)	11.1 (U: 41, D: 63)	11.3 (U: 41, D: 65)	Primary metabolic process, protein modification process, carbohydrate biosynthetic process	Nucleic acid metabolic process, cellular nitrogen compound metabolic process, heterocycle metabolic process

4 Notes

1. The NCBI GEO is a repository for array- and sequence-based data. The database can be searched using keywords, e.g., "drought" and search results can be further filtered based on various options such as "organism," "type of study," "platform," and so on. The current dataset was selected on the basis of the stress condition, "drought," organism, i.e., "*Brachypodium distachyon*," the study type, i.e., "expression profiling by array," and platform "Affymetrix."

2. For Affymetrix platform, one can start the analysis from scratch by downloading the raw .CEL files from GEO. The "simpleaffy" package can be used to read the files along with the phenotypic data and the CDF file. The CDF file is usually downloaded by BioConductor automatically while executing the steps. However, if the file is not present in the BioConductor, especially in case of custom arrays, it has to be obtained from GEO, Affymetrix or elsewhere and packages like "makecdfenv" in BioConductor has to be used before one can proceed with the normalization steps. Otherwise, the SOFT files with preprocessed data can be used as demonstrated in the manuscript. #command.
   The following R commands are given in case .CEL files are used:

```
>library(simpleaffy)
>raw.data <- read.affy()
>data_rma <- call.exprs(raw.data,"rma") #data_rma contains
the normalized, log-transformed expression matrix
```

3. The user is advised to check the prerequisites for WGCNA and the version of R. Detailed instructions are given in [ ]
4. While reading the expression matrix, it is important to see if there are too many missing values and "gsg$allOK" should return TRUE as elaborated in Subheading 3.2.1. Else, the following R commands should be given to remove those genes/samples:

```
>if (!gsg$allOK)
{
#Optionally, print the gene and sample names that were
removed:
if (sum(!gsg$goodGenes)>0)
printFlush(paste("Removing genes:", pastenames(datExpr0)[!
gsg$goodGenes], collapse = ", "));
if (sum(!gsg$goodSamples)>0)
printFlush(paste("Removing samples:", pasterownames(da-
tExpr0)[!gsg$goodSamples], collapse = ", "));
# Remove the offending genes and samples from the data:
datExpr0 = datExpr0[gsg$goodSamples, gsg$goodGenes]
}
```

5. For Identification of differentially expressed genes, the choice of fold-change and *p*-value cutoffs can vary among the users. Accordingly, the number of DEGs will differ. The user is advised to see the fold-change distribution across the genes and then decide on the thresholds.

6. Network construction using WGCNA involves several steps from data loading, data checking, selection of soft-thresholds, module detection, computing centrality measures, functional enrichment and finally exporting the networks. It will be convenient for the users if the commands for each of these sections are written in separate r scripts so that rerunning any particular task will become easier. Moreover, saving the R sessions after each task is advisable.
7. The functions to be used for network construction should be decided by the number of genes/probes given as the input and the memory of the computer as elaborated in Subheading 3.2.3.
8. The parameters used in the module detection step have to be tweaked by the user over several iterations. The parameter "deepSplit" and the dendrogram cut-height can be varied until a desired number of modules are obtained.
9. The cutoff for the edge-weights of the networks exported has been kept as 0.02. However, the user may want to increase this cutoff in case loading to Cytoscape becomes difficult.
10. If the annotation package is available in BioConductor for a given organism, one can perform enrichment analysis in WGCNA by giving the following R commands:

```
>GOenr = GOenrichmentAnalysis(moduleColors, allLLIDs, organism = "mouse", nBestP = 10); # returns the 10 best terms for each module
>tab = GOenr$bestPTerms[[4]]$enrichment
>names(tab) #extract the names of the columns
> write.table(tab, file = "GOEnrichmentTable.csv", sep = ",", quote = TRUE, row.names = FALSE) # save the table in a file
```

References

1. Massa AN, Childs KL, Buell CR (2013) Abiotic and biotic stress responses in group Phureja DM1-3 516 R44 as measured through whole transcriptome sequencing. Plant Genome 6:3
2. Bassel GW, Lan H, Glaab E et al (2011) Genome-wide network model capturing seed germination reveals coordinated regulation of plant cellular phase transitions. Proc Natl Acad Sci U S A 108:9709–9714
3. Sircar S, Parekh N (2015) Functional characterization of drought-responsive modules and genes in Oryza sativa: a network-based approach. Front Genet 6:256
4. Appel HM, Fescemyer H, Ehlting J et al (2014) Transcriptional responses of Arabidopsis thaliana to chewing and sucking insect herbivores. Plant Microbe Interact 5:565
5. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559
6. Verelst W, Bertolini E, De Bodt S et al (2013) Molecular and physiological analysis of growth-limiting drought stress in Brachypodium distachyon leaves. Mol Plant 6:311–322
7. Wilson CL, Miller CJ (2005) Simpleaffy: a BioConductor package for affymetrix quality control and data analysis. Bioinformatics (Oxford) 21:3683–3685

8. Bolser DM, Kerhornou A, Walts B et al (2015) Triticeae resources in ensembl plants. Plant Cell Physiol 56:e3
9. Smoot ME, Ono K, Ruscheinski J et al (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics (Oxford) 27:431–432
10. Goodstein DM, Shu S, Howson R et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40: D1178–D1186
11. Du Z, Zhou X, Ling Y et al (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res 38:W64–W70
12. Monaco MK, Stein J, Naithani S et al (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Res 42: D1193–D1199

Chapter 17

Whole Genome DNA Methylation Analysis Using Next-Generation Sequencing (BS-seq)

I-Hsuan Lin

Abstract

Plant methylation is widely evident and has played crucial roles ranging in defining the epi-genome variations during abiotic and biotic stress. Variations in epi-genomic level has observed not only in the symmetrical as well as the non-symmetrical sequences. Plethora of these epi-genomic variations have been widely also demonstrated at the flowering, tissue-specific, and also at developmental stages revealing a strong association of the observed epi-alleles to the physiological state. In the present chapter, epi-genomic analysis of the *s* has been described with functional workflow and illustrated methodology.

Key words *Brachypodium distachyon*, BS-seq, Methylation, Epigenomics

1 Introduction

Plant methylation has been widely studied and has played an important and intricate role in understanding the adaptation genetics. Observations of these cytosine methylation as epigenetic marks are not only observed at the symmetric but also at the non-symmetric DNA sequence level. Among the methylation levels, transposons have been shown to be highly methylated, which plays a major role in silencing the transposons. Epigenetic methylation levels have not only been described widely regulated at the developmental stages but also during the abiotic and biotic stress thus elucidating the cross talks between the epi-genetic imprinting and stress adaptation. Relative links between the methylation and smallRNAs have been widely illustrated in plants with base excision repair pathway been the widely studied pathway in plants. In this protocol chapter, using *Brachypodium distachyon* as a model plant, a complete

Electronic supplementary material: The online version of this chapter (doi:10.1007/978-1-4939-7278-4_17) contains supplementary material, which is available to authorized users.

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_17, © Springer Science+Business Media LLC 2018

protocol for understanding the methylation signals and processing of the methylation data has been provided using flowering methylomes.

2 Materials and Requirements

In this chapter, processing and visualization of *Brachypodium distachyon* BS-seq data that are available for download from the NCBI BioProject website (accession PRJNA182267 [1]) have been described. The MethPipe (DNA Methylation Data Analysis Pipeline) suite will be used to perform methylome construction [2]. I will introduce several R packages that can be used for downstream visualization, interpretation and comparison of samples in this BS-seq dataset.

2.1 Hardware

The software used in this book chapter is designed to run in a UNIX-like operating environment. The processing of NGS data can be computationally intensive and demand large memory and I/O capabilities, hence utilizing of cluster computing with common schedulers such as LSF, PBS, and SGE is highly recommended. Ideally, individual nodes will have 64GB of memory for processing and at least 1 TB disk space for storing raw data and intermediate files.

2.2 Software

Readers are assumed to be familiar with UNIX-like operating environment and able to build and install the required software from source files. Tables 1 and 2 list the software and R packages that will be used in this chapter.

2.3 Sequence Data

Table 3 shows a list of required sequence files in this chapter. The JGI v3.0 assembly of *B. distachyon* Bd21 is downloaded from Phytozome database [3]. Phytozome database v11 released on 12 Jan 2016 is the latest version at the time of writing. Registration and login is required to download bulk data from the Phytozome database. The tandem repeat and transposable element annotation

Table 1
List of required software packages

Name	Version	URL
SRA Toolkit	2.5.7	https://github.com/ncbi/sra-tools/wiki/Downloads
WALT	0.7	https://github.com/smithlabcode/walt/releases
MethPipe	3.4.2	https://github.com/smithlabcode/methpipe/releases
BedTools	2.25.0	https://github.com/arq5x/bedtools2
R	3.2.4	https://www.r-project.org/

Table 2
List of required R packages

Name	URL	Repository
devtools	https://cran.r-project.org/package=devtools	CRAN
data.table	https://cran.r-project.org/package=data.table	CRAN
reshape2	https://cran.r-project.org/package=reshape2	CRAN
ggplot2	https://cran.r-project.org/package=ggplot2	CRAN
GenomicRanges	http://bioconductor.org/packages/GenomicRanges/	Bioconductor
IRanges	http://bioconductor.org/packages/IRanges/	Bioconductor
rtracklayer	https://bioconductor.org/packages/rtracklayer/	Bioconductor
methylKit	https://github.com/al2na/methylKit	GitHub
EnrichedHeatmap	https://github.com/jokergoo/EnrichedHeatmap	GitHub
circlize	https://github.com/jokergoo/circlize	GitHub
methylPipe	https://github.com/ycl6/methylPipe	GitHub
BSgenome.Bdistachyon.JGI.Bd3.1	https://github.com/ycl6/BSgenome.Bdistachyon.JGI.Bd3.1	GitHub
TxDb.Bdistachyon.JGI.Bd3.1.geneexons	https://github.com/ycl6/TxDb.Bdistachyon.JGI.Bd3.1.geneexons	GitHub

files is retrieved from the PGSB/MIPS FTP site. The chloroplast genome can be obtained from the NCBI Nucleotide database under the accession EU325680. The FASTA sequence is saved into a file named chloroplast.fa. An excerpt of the FASTA file is shown below:

>EU325680.1 Brachypodium distachyon cultivar Bd21 chloroplast, complete genome.

AGGGAAATACCCCAATATCTTGGGTAGGAACAAGA-
TATTGGGTATTTCTCGCTTTTCTTTTCTTCAAAAATTCT-
TATATGTTAGCGGAAAAACCTTATCCATTAATAGCGG-
GAACTTCAAGAGCAGCTAGATCTAGAGGGAAGTTGT-
GAGCATTACGTTCGTGCATTACTTCCATACCAAGATTAG-
CACGGTTGATGATATCAGCCCAAGTATTAATAACGC-
GACCTTGACTATCAACTACAGATTGGTTGAAATTGAATC-
CATTTAGGTTGAACGCCATAGTACTAATACCTAAAG-
CAGTGAACCAGATTCCTACTACAGGCCAAGCAGCCAA-
GAAGAAGTGTAAAGAA

The BS-seq raw data of *B. distachyon* Bd21 is deposited in the NCBI Sequence Read Archive (SRA) in a special compressed format. The fastq-dump program from the SRA Toolkit is used to retrieve, decompress and convert the SRA files into FASTQ files.

Table 3
List of F required ASTA and SRA files

Name	Description	Provider	Size
Bdistachyon_314_v3.0.fa.gz	Genome assembly	Phytozome	78 MB
chloroplast.fa	Chloroplast genome	NCBI	134 kB
MIPS_Bd_tandemRepeats_01-04-2009.gff3.gz	Tandem repeats	PGSB/ MIPS	2315 kB
MIPS_Bd_Transposons_v2.2_16-07-2009.gff3. gz	Transposons	PGSB/ MIPS	574 kB
SRR628921.sra	Leaf rep 1	NCBI SRA	12 GB
SRR629088.sra	Leaf rep 2	NCBI SRA	11 GB
SRR629207.sra	Leaf rep 3	NCBI SRA	11 GB
SRR629437.sra	Immature floral buds rep 1	NCBI SRA	11 GB
SRR629438.sra	Immature floral buds rep 2	NCBI SRA	9.9 GB
SRR629439.sra	Immature floral buds rep 3	NCBI SRA	9.5 GB

The sequences from Read 1 and Read 2 are saved into separate files with the –split-files option. The FASTQ files are placed in the current working directory (".") and –outdir option can be used to specify a different location. Execute fastq-dump to obtain the six pairs of FASTQ files:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
fastq-dump -split-files ${id}
done
```

3 Methods

3.1 Generate Genome Index

The *B. distachyon* Bd21 genomic and chloroplast sequence files are merged into a single file by running:

```
$ zcat Bdistachyon_314_v3.0.fa.gz | cat - chloroplast.fa >
Bdistachyon_314_v3.0_ch.fa
```

The mapping of BS-seq reads is achieved by using WALT (Wildcard ALignment Tool). Before mapping, the makedb program from WALT is used to produce the index using the *B. distachyon* Bd21 genome assembly as follow:

```
$ makedb -c Bdistachyon_314_v3.0_ch.fa -o Bdista-
chyon_314_v3.0_ch.dbindex
```

After the process is completed, use the ls command to view the five newly created files:

```
$ ls -lh Bdistachyon_314_v3.0_ch.dbindex* | awk '{ print $5,$8 }'

209 Bdistachyon_314_v3.0_ch.dbindex
1.4G Bdistachyon_314_v3.0_ch.dbindex_CT00
1.4G Bdistachyon_314_v3.0_ch.dbindex_CT01
1.4G Bdistachyon_314_v3.0_ch.dbindex_GA10
1.4G Bdistachyon_314_v3.0_ch.dbindex_GA11
```

3.2 Map BS-seq Reads with WALT

As described in the MethPipe manual [4], in paired-end sequencing the reads from Read 1 file (5′ reads or mate 1) are often T-rich, whereas reads from Read 2 file (3′ reads or mate 2) are often A-rich. The walt program assumes both files contain equal number of reads and are in the same order. Execute the following command to map the paired read files of SRR628921:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
   walt -i Bdistachyon_314_v3.0_ch.dbindex -1 ${id}_1.fastq -2
${id}_2.fastq -o ${id}.mr -C AGATCGGAAGAGC -m 5 -u -a -t 46
done
```

The required options are -i (index file), -1 (Read 1 file), -2 (Read 2 file) and -o (output file). The example also specifies five other options: (1) clip the Illumina standard adapters (AGATCGGAAGAGC) with -C, (2) allow at most five mismatches with -m, (3) output unmapped reads into a separated file with -u, (4) randomly output one mapped position for ambiguous reads into a separated file with -a, and (5) use 46 threads for mapping specified with -t. For future reference when working with single-end sequencing data, the command to map single-end BS-seq reads is as follow:

```
$ walt -i <index file> -r <read file> -o <output file>
[options]
```

WALT can output alignment in SAM (.sam) or MR (.mr) format. Here, the default MR format was chosen. If SAM output was chosen or if other raw read aligners (such as BSSeeker, Bismark or BSMAP) were used, the to-mr program from MethPipe can be used to convert SAM/BAM files to the MR format suitable for downstream processing. For example:

```
$ to-mr -o SRR628921.mr -m general SRR628921.bam
$ to-mr -o SRR628921.mr -m bs_seeker SRR628921.bs_seeker.bam
$ to-mr -o SRR628921.mr -m bismark SRR628921.bismark.bam
$ to-mr -o SRR628921.mr -m bsmap SRR628921.bsmap.bam
```

3.3 Remove Duplicated Reads

Reads that have identical sequence and mapped to the same genomic location, i.e. same chromosome, start and end positions and strand, are most likely the result of PCR amplification and should

be removed before calculating DNA methylation. By executing the duplicate-remover program from MethPipe, it will randomly select one read from a set of the duplicated reads as representative and remove the remaining duplicated reads.

The MR file has to be sorted in the following order: chromosome, start, end, and strand before executing the duplicate-remover program. The GNU core utilities sort is used to perform sorting. In the below example, 46 sort instances were ran concurrently (-parallel option) and allowing a maximum of 120G memory (-S option).

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
    LC_ALL=C sort -T . -parallel=46 -S 120G -k 1,1 -k 2,2n -k
3,3n -k 6,6 -o ${id}.sorted_start ${id}.mr
done
```

Next, executing the duplicate-remover program to remove duplicate reads:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
     duplicate-remover -v -S ${id}.dremove_stat.txt -o ${id}.
dremove ${id}.sorted_start
done
```

In the above command, −v option enables the verbose mode and -S option specifies the statistics output file. Additionally, −s option can be used to allow the program to take sequence information into account beside genomic location, i.e. reads of identical genomic location but different CpG occurrences (i.e. different methylation pattern) are not considered duplicates. The -s -A options will consider all cytosines.

3.4 Estimate Bisulfite Conversion Rate

To estimate the bisulfite conversion rate in plant genome, the chloroplast genome is used as a control whereby the cytosines in the chloroplast genome are believed to be unmethylated. Given the reads that are mapped to the chloroplast genome, the bsrate program from MethPipe identifies the cytosines that are presumed unmethylated, then compute the ratio of C to (C + T) at these positions. The ratio will be close to 1 if all cytosines are converted in the bisulfite reaction.

From the mapping result, the reads mapped to the chloroplast genome is obtained using grep:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
```

Table 4
Overall conversion rates of the six samples

Sample ID	Conversion rate
SRR628921	0.996537
SRR629088	0.995993
SRR629207	0.996532
SRR629437	0.996377
SRR629438	0.996608
SRR629439	0.996112

```
    LC_ALL=C grep ^EU325680.1 ${id}.dremove > ${id}.chlor-
oplast
done
```

Then, the conversion rate is estimated using bsrate. The -N option allow the program to count all cytosines, including CpG sites. The -c option specifies the file containing the chloroplast genomic sequence. The estimated conversion rates of the six BS-seq samples are listed in Table 4.

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
     bsrate -N -v -c chloroplast.fa -o ${id}.bsrate ${id}.
chloroplast
done
```

3.5 Compute Methylation Levels and Statistics

The methcounts program from MethPipe is used to estimate the DNA methylation level for every cytosine site at single base resolution. Before computing DNA methylation levels at cytosines. It is necessary to sort the file containing the mapped reads and with duplicated reads removed (*see* Subheading 3.3). Execute the below command to call methylation at each cytosine of all samples:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
     methcounts -c Bdistachyon_314_v3.0_ch.fa -v -o ${id}.
methcount ${id}.dremove
done
```

The output from methcounts can be fed into the levels program from MethPipe to compute methylation level statistics as follow:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
```

```
SRR629439
do
    levels -o ${id}.levels ${id}.methcount
done
```

3.6 Analyze DNA Methylation with R/Bioconductor Package methylKit

The methylKit package was first developed and published in 2012 [5]. Since its release, it has been widely used to analyze base-resolution methylation and hydroxymethylation sequencing data. Therefore, I will give a thorough step-by-step guide to all its functions and features in the following subsections using methylation call files (*.methcount) generated by MethPipe.

3.6.1 Installation of R Packages

In R environment, install the latest version of devtools, data.table and ggplot2 from CRAN by running:

```
> install.packages(c("data.table","devtools","ggplot2"))
```

Install the latest version of GenomicRanges and IRanges from bioconductor by running:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite(c("GenomicRanges","IRanges"))
```

Then, the install_github function of devtools is used to install the required (and dependent) packages from GitHub:

```
devtools::install_github("al2na/methylKit")
devtools::install_github("ycl6/TxDb.Bdistachyon.JGI.Bd3.1.
geneexons")
```

3.6.2 Reading of Methylation Call Files

Before using methylKit in R, the methylation call files generated by MethPipe is converted to the format that can be recognized by methylKit. A typical input file for methylKit looks like this:

chrBase	chr	base	strand	coverage	freqC	freqT
Bd1.259	Bd1	259	F	2	100	0
Bd1.260	Bd1	260	R	58	94.83	5.17
Bd1.353	Bd1	353	F	9	100	0
Bd1.354	Bd1	354	R	18	94.44	5.56
Bd1.366	Bd1	366	F	11	81.82	18.18

The input file requires positions to have at least one read coverage and has to contain the proportion of reads correspond to cytosine and thymine at each base. The three awk commands below extract positions corresponding to the three cytosine contexts, i.e. CpG, CHH and CHG, respectively, to convert *.methcount file to the required format.

```
mkdir methylKit methylKit/CpG methylKit/CHG methylKit/CHH
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438
SRR629439
do
    awk -F $'\t' 'BEGIN { OFS=FS } { if($6 > 0 && ($4 == "CpG"
|| $4 == "CpGx")) { std = "F"; if($3 == "-") std = "R"; C =
sprintf("%.0f",$5*$6); T = $6-C; print $1"."$2,$1,$2,std,$6,
sprintf("%.2f",C/$6*100),sprintf("%.2f",T/$6*100) } }' ${id}.
methcount > methylKit/CpG/${id}.txt
    awk -F $'\t' 'BEGIN { OFS=FS } { if($6 > 0 && ($4 == "CHH"
|| $4 == "CHHx")) { std = "F"; if($3 == "-") std = "R"; C =
sprintf("%.0f",$5*$6); T = $6-C; print $1"."$2,$1,$2,std,$6,
sprintf("%.2f",C/$6*100),sprintf("%.2f",T/$6*100) } }' ${id}.
methcount > methylKit/CHH/${id}.txt
    awk -F $'\t' 'BEGIN { OFS=FS } { if($6 > 0 && ($4 == "CCG"
|| $4 == "CCGx" || $4 == "CXG" || $4 == "CXGx")) { std = "F";
if($3 == "-") std = "R"; C = sprintf("%.0f",$5*$6); T = $6-C;
print  $1"."$2,$1,$2,std,$6,sprintf("%.2f",C/$6*100),sprintf
("%.2f",T/$6*100) } }' ${id}.methcount > methylKit/CHG/${id}.
txt
done
```

In R environment, the read function is used to read a list of file paths leading to the input files. A methylRawList object containing six methylRaw objects is created in the process. The sample.id vector defines the sample names and assembly string defines the genome assembly. The treatment vector is used to define the sample (s) correspond to control (0) and test (1) sets. The context string defines the sequence context of the target cytosine, such as CpG, CpH, or CHH.

```
library(methylKit)

# Paths
cpg.files = as.list(list.files(file.path("methylKit/CpG"),
full.names=TRUE))
chg.files = as.list(list.files(file.path("methylKit/CHG"),
full.names=TRUE))
chh.files = as.list(list.files(file.path("methylKit/CHH"),
full.names=TRUE))

# Read files
mobj.cpg = read(cpg.files, sample.id=list("leaf1","lea-
f2","leaf3","spike1","spike2","spike3"), assembly="Bdistach-
yon", treatment=c(0,0,0,1,1,1), context="CpG")

mobj.chg = read(chg.files, sample.id=list("leaf1","lea-
f2","leaf3","spike1","spike2","spike3"), assembly="Bdistach-
yon", treatment=c(0,0,0,1,1,1), context="CHG")

mobj.chh = read(chh.files, sample.id=list("leaf1","lea-
```

```
f2","leaf3","spike1","spike2","spike3"), assembly="Bdistachyon", treatment=c(0,0,0,1,1,1), context="CHH")
```

The methylRawList object is filtered to discards cytosine positions that have less than 10x coverage in each sample. Bases that have more than 99.9th percentile of coverage (i.e. exceptionally high coverage that could be PCR duplication bias) are also discarded.

```
filt.mobj.cpg = filterByCoverage(mobj.cpg, lo.count=10, hi.perc=99.9)
filt.mobj.chg = filterByCoverage(mobj.chg, lo.count=10, hi.perc=99.9)
filt.mobj.chh = filterByCoverage(mobj.chh, lo.count=10, hi.perc=99.9)
```

The construction of the methylRawList object is a lengthy process for such a large dataset. Therefore It is advisable to save the objects at this time if you wish to perform the analysis at a later time.

```
> save(mobj.cpg, filt.mobj.cpg, mobj.chg, filt.mobj.chg, mobj.chh, filt.mobj.chh, file="methylRawList.RData")
```

To load the object into R in a new session, execute:

```
> load("methylRawList.RData")
```

3.6.3 Basic Statistics About the Methylation Data

The getMethylationStats and getCoverageStats functions can print the percentage methylation distribution and read coverage per base statistics to the R console, as well as generate histograms to visualize these statistics. The calculation of CpG methylation statistics are presented below, please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts. The sink function sends R output, including the percentage of bases that pass the filtering process to a text file. The histograms generated are saved as PNG files (*see* Figs. 1, 2, and 3).

```
sink("methylKit.cpg.statistics.txt")
for (i in c(1:length(mobj.cpg)))

{sample.id = mobj.cpg[[i]]@sample.id
pPass = nrow(filt.mobj.cpg[[i]])/nrow(mobj.cpg[[i]])
print(paste(sample.id, ":", sprintf("%.2f",pPass*100),"%"))

    # % methylation distribution
    getMethylationStats(filt.mobj.cpg[[i]], plot=F, both.strands=F)

    # Plot
    png(paste(sample.id,".cpg.methylation.png", sep=""))
    getMethylationStats(filt.mobj.cpg[[i]], plot=T, both.
```

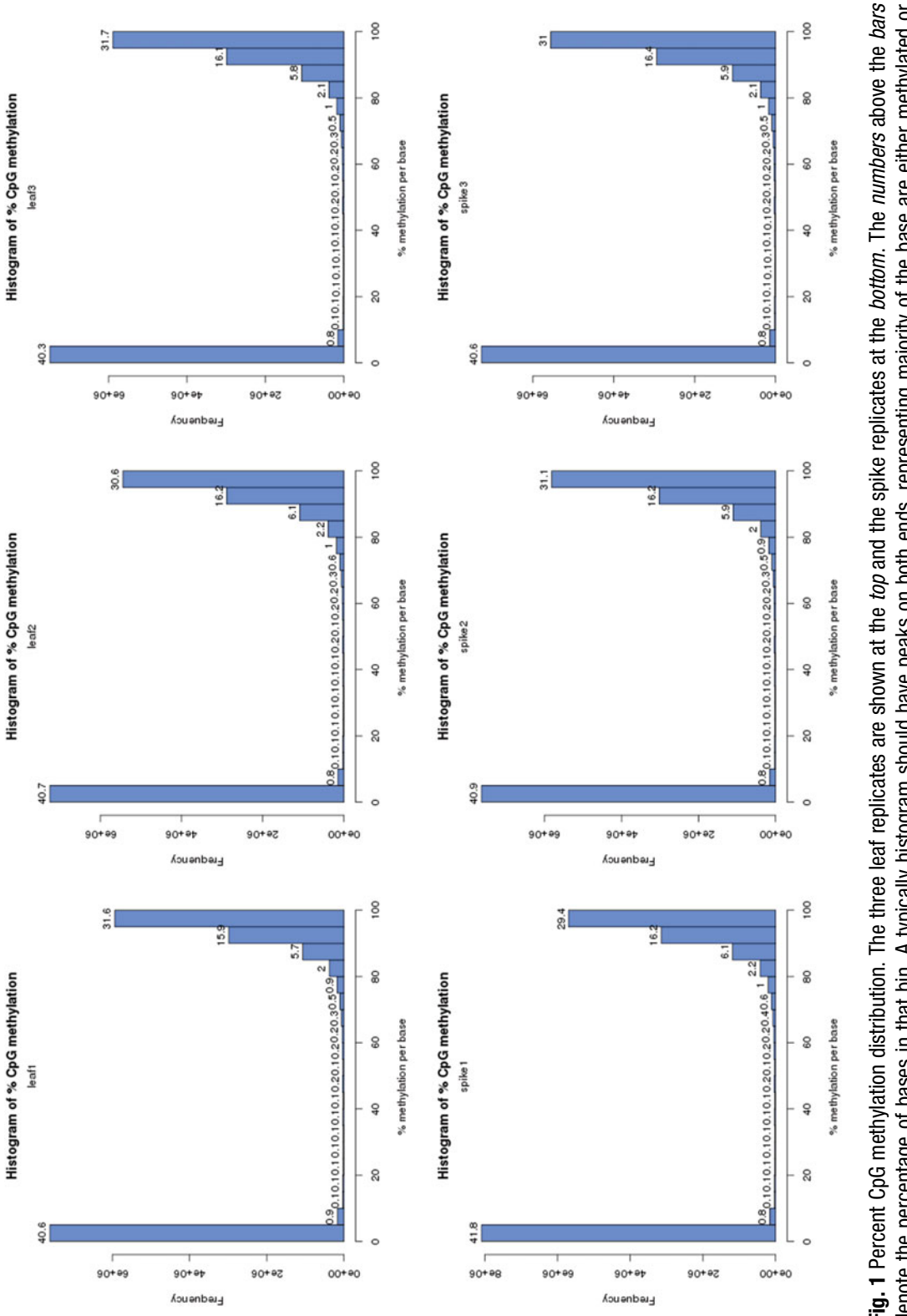

Fig. 1 Percent CpG methylation distribution. The three leaf replicates are shown at the *top* and the spike replicates at the *bottom*. The *numbers* above the *bars* denote the percentage of bases in that bin. A typically histogram should have peaks on both ends, representing majority of the base are either methylated or unmethylated

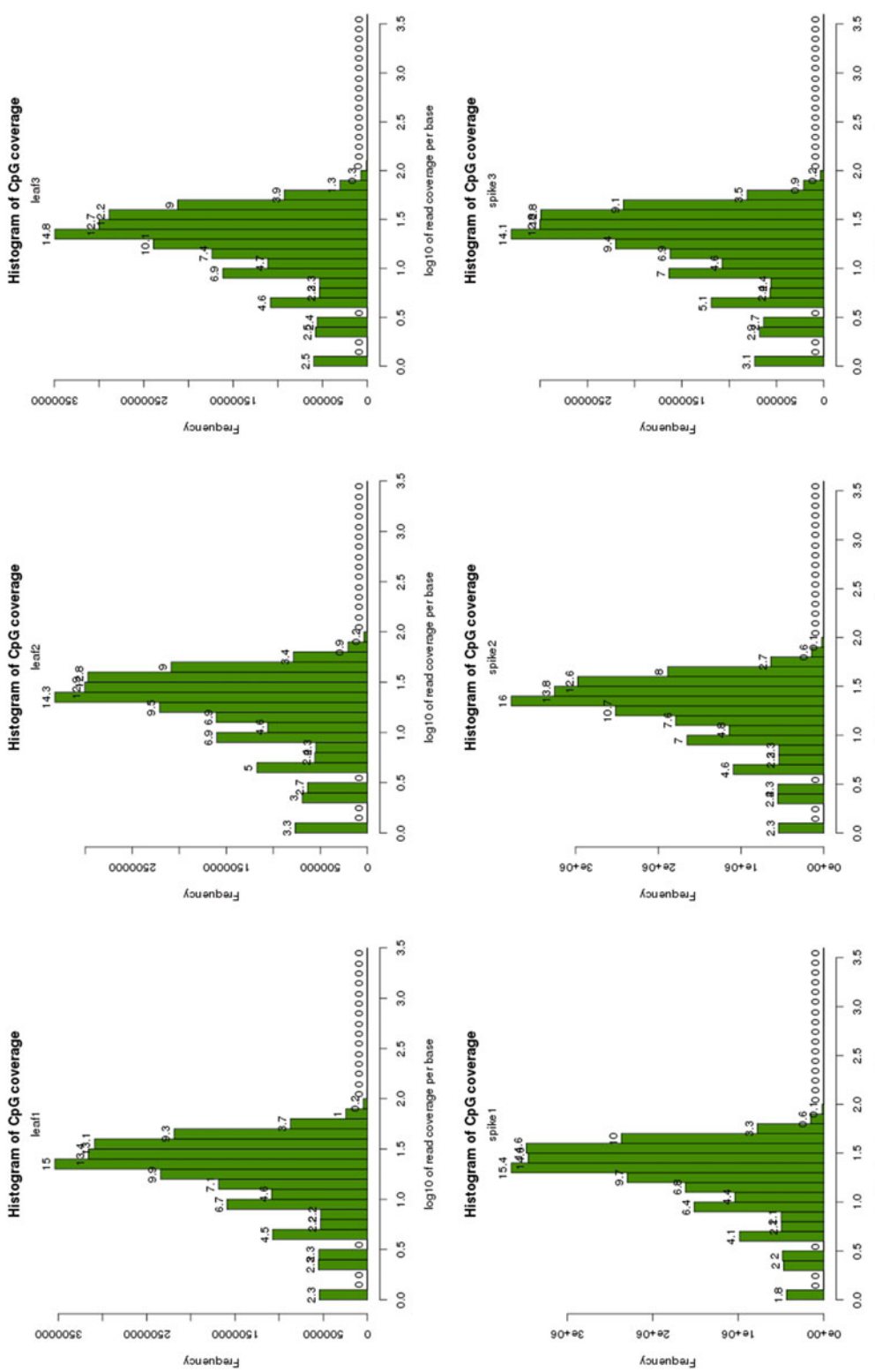

Fig. 2 Read coverage per base of CpG sites before applying filterByCoverage. The three leaf replicates are shown at the *top* and the spike replicates at the *bottom*. The *numbers* above the *bars* denote the percentage of bases in that bin. Experiments that suffers from PCR duplication bias will show a second peak towards the right. In these examples, the samples do not suffer from this bias

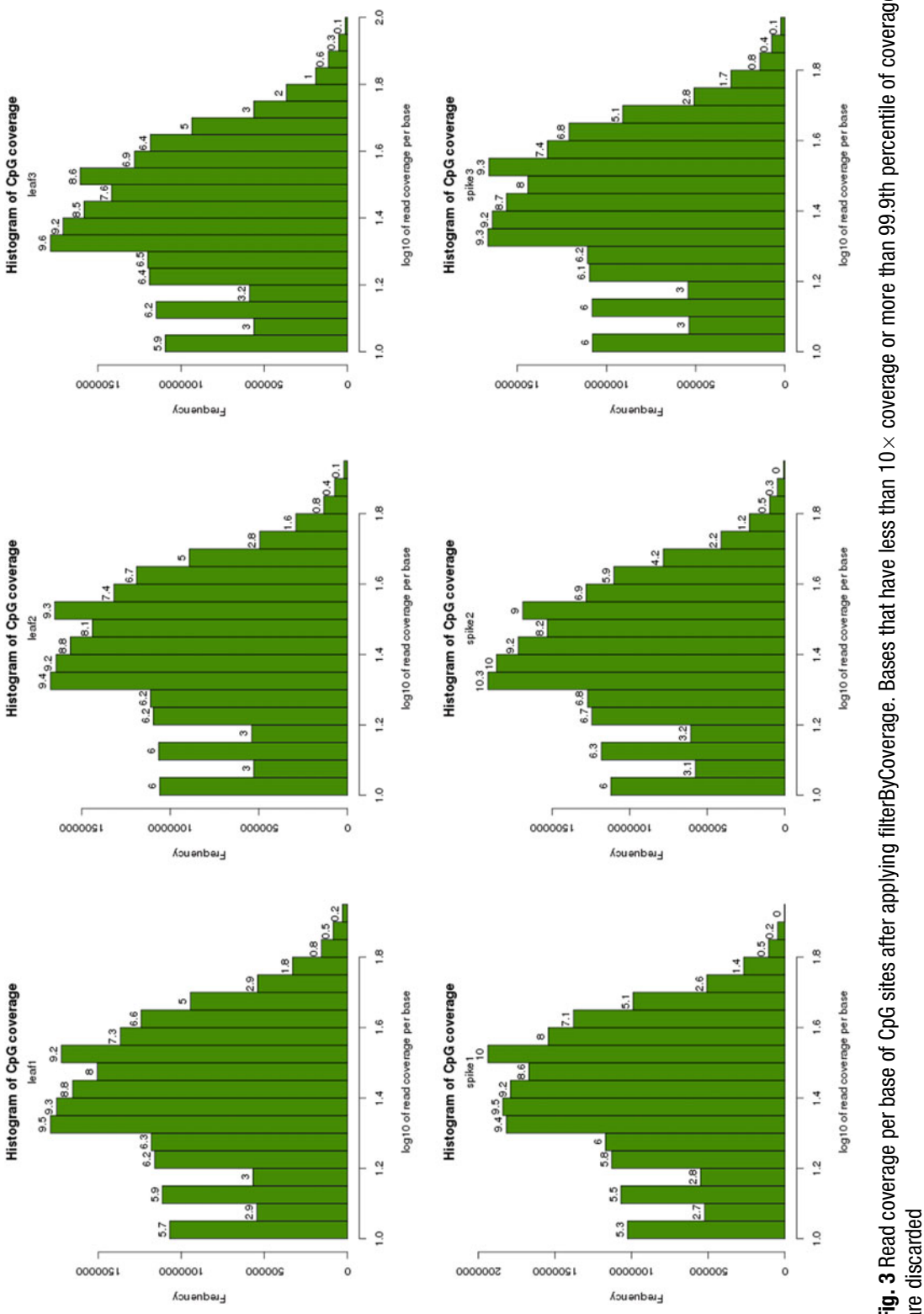

Fig. 3 Read coverage per base of CpG sites after applying filterByCoverage. Bases that have less than 10× coverage or more than 99.9th percentile of coverage are discarded

```
strands=F)
    dev.off()

    # read coverage per base information
    getCoverageStats(filt.mobj.cpg[[i]], plot=F, both.
strands=F)

    # Plot coverage before filtering, mobj.cpg
    png(paste(sample.id,".cpg.ori-coverage.png", sep=""))
    getCoverageStats(mobj.cpg[[i]], plot=T, both.strands=F)
    dev.off()

    # Plot coverage after filtering, filt.mobj.cpg
    png(paste(sample.id,".cpg.coverage.png", sep=""))
    getCoverageStats(filt.mobj.cpg[[i]], plot=T, both.
strands=F)
    dev.off()
}
sink()
```

3.6.4 Comparative Analysis of Methylome Profiles

Before comparing the methylation profiles between samples, the unite function is used to merge and retain only bases that are covered in all samples. There are two arguments that controls how the data is processed. The destrand=TRUE (default is FALSE) will merge reads on both strands to provide higher coverage, however this is appropriate only when analyzing methylation of the CpG context. Also, it works only on base-resolution methylome data. The min.per.group defines the minimum number of replicates per sample group needed to cover a region or base. For example, min.per.group=2L will keep CpG sites covered in at least two replicates per group.

```
meth.cpg = unite(filt.mobj.cpg) # CpG sites covered in all
samples.
meth.chg = unite(filt.mobj.chg) # CHG sites covered in all
samples.
meth.chh = unite(filt.mobj.chh) # CHH sites covered in all
samples.
```

Retrieve the dimension with dim function:

```
> dim(meth.cpg)
[1] 15783683       22
```

Use head function to view part of the content of the object:

```
> head(meth.cpg,3)
methylBase object with 3 rows
_____________

   chr start   end strand coverage1 numCs1 numTs1 coverage2
numCs2 numTs2
```

```
1 Bd1   260  260      -        58     55      3        66     59
7
2 Bd1   354  354      -        18     17      1        22     22
0
3 Bd1  1092 1092      -        15     14      1        11     10
1
   coverage3 numCs3 numTs3 coverage4 numCs4 numTs4 coverage5
numCs5 numTs5
1        86     82      4        60     56      4        54     53
1
2        39     39      0        13     13      0        13     13
0
3        15     14      1        18     16      2        23     21
2
  coverage6 numCs6 numTs6
1        73     68      5
2        17     16      1
3        11     11      0
_____________
sample.ids: leaf1 leaf2 leaf3 spike1 spike2 spike3
destranded FALSE
assembly: Bdistachyon
context: CpG
treament: 0 0 0 1 1 1
resolution: base
```

Save the three objects:

```
save(meth.cpg, meth.chg, meth.chh, file="methylBase.RData")
```

After obtaining the merged methylation data, pairwise correlation, sample clustering and principal component analysis can be performed. The getCorrelation function prints the Pearson matrix of correlation coefficients, and the method argument allows other correlation methods such as Spearman and Kendall to be computed. The plot setting controls if scatterPlot is produced at the same time. Figure 4 shows the graphical output of getCorrelation of the CpG context, please refer to Supplementary Files 1 and 2 for CHG and CHH graphical outputs.

```
sink("methylKit.cpg.correlation.txt")
png("methylKit.cpg.correlation.png")
getCorrelation(meth.cpg, plot=TRUE)
dev.off()
sink()
```

The clusterSamples function uses hclust function to performs hierarchical clustering samples based on the similarity of their methylation profiles and produce a dendrogram. The dist argument sets the distance measure to be used and method defines the

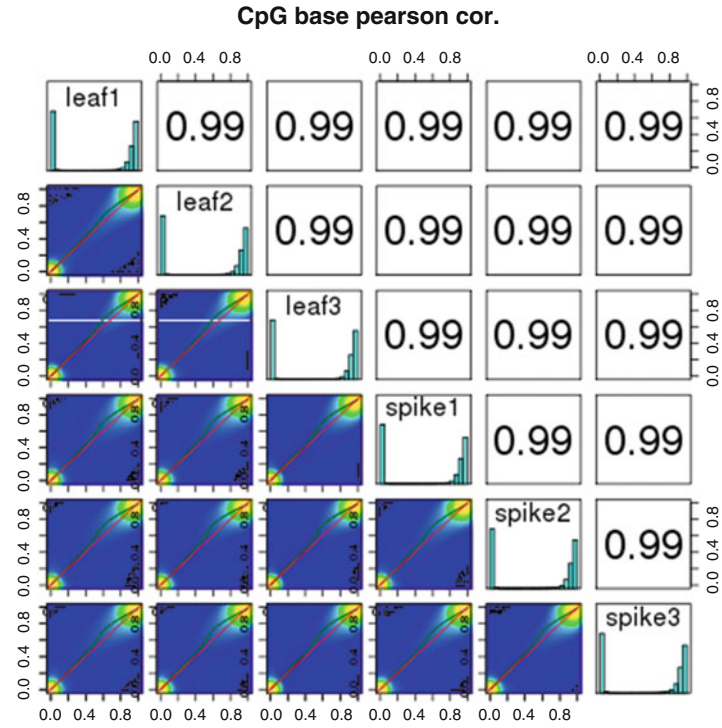

Fig. 4 CpG methylation correlation plot of the six samples

agglomeration method to be used. When plot=FALSE, the function will return a dendrogram object. Figure 5 shows the graphical output of clusterSamples of the CpG context, please refer to Supplementary Files 1 and 2 for CHG and CHH graphical outputs.

```
sink("methylKit.cpg.methcluster.txt")
png("methylKit.cpg.methcluster.png")
clusterSamples(meth.cpg, dist="correlation", method="ward",
plot=TRUE)
dev.off()
sink()
```

Additionally, the PCASamples function performs principal component (PC) analysis using the prcomp function. When screeplot=FALSE, a scatter plot of PC1 and PC2 for the PC model is produced instead of a scree plot to illustrate the rate of change in the magnitude of the eigenvalues for the PC (*see* Figs. 6 and 7).

```
# screeplot=FALSE
png("methylKit.cpg.pca_scatterplot.png")
PCASamples(meth.cpg, screeplot=FALSE)
dev.off()

# screeplot=TRUE
png("methylKit.cpg.pca_screeplot.png")
PCASamples(meth.cpg, screeplot=TRUE)
dev.off()
```

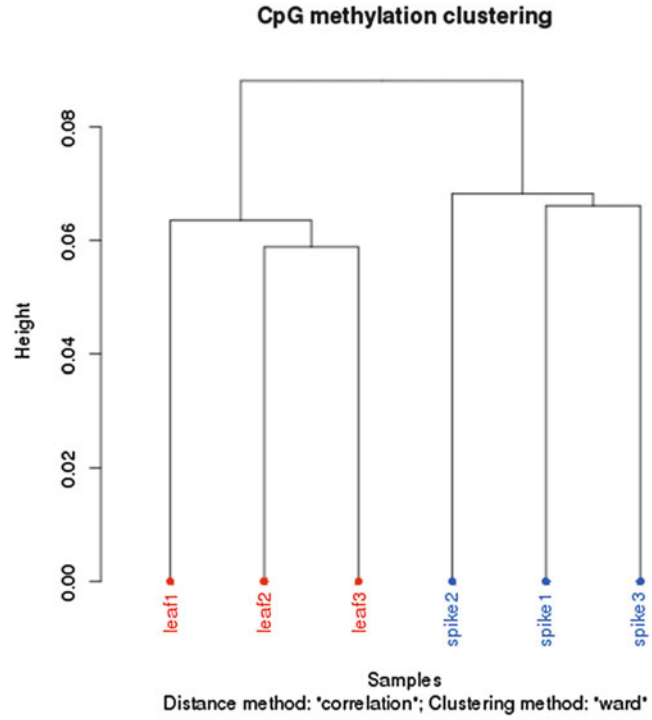

Fig. 5 Dendrogram showing the hierarchical clustering of replicates of the two plant tissues

3.6.5 Identify Differentially Methylated Bases

The calculateDiffMeth function is called to calculate differential methylation and it produces a methylDiff object. When there are multiple samples per group the function uses a logistic regression test to calculate differential methylation, whereas Fisher's Exact test is used when there is one sample per group. The Sliding Linear Model (SLIM) method is applied to correct for multiple hypothesis testing by default (slim=TRUE), If FALSE, the function uses the Benjamini-Hochberg (BH) method for P-value correction. This function can be run in parallel on multiple cores by setting the num.cores parameter, for example num.cores=2.

To use the function, run:

```
mDiff.cpg = calculateDiffMeth(meth.cpg)
mDiff.chg = calculateDiffMeth(meth.chg)
mDiff.chh = calculateDiffMeth(meth.chh)
```

Use head function to view part of the content of the object:

```
> head(mDiff.cpg)
Object of class "methylDiff"
  chr start  end strand    pvalue    qvalue   meth.diff
1 Bd1   260  260      - 0.5808918 0.9211514  1.31907308
2 Bd1   354  354      - 0.6667881 0.9211514 -1.05975861
```

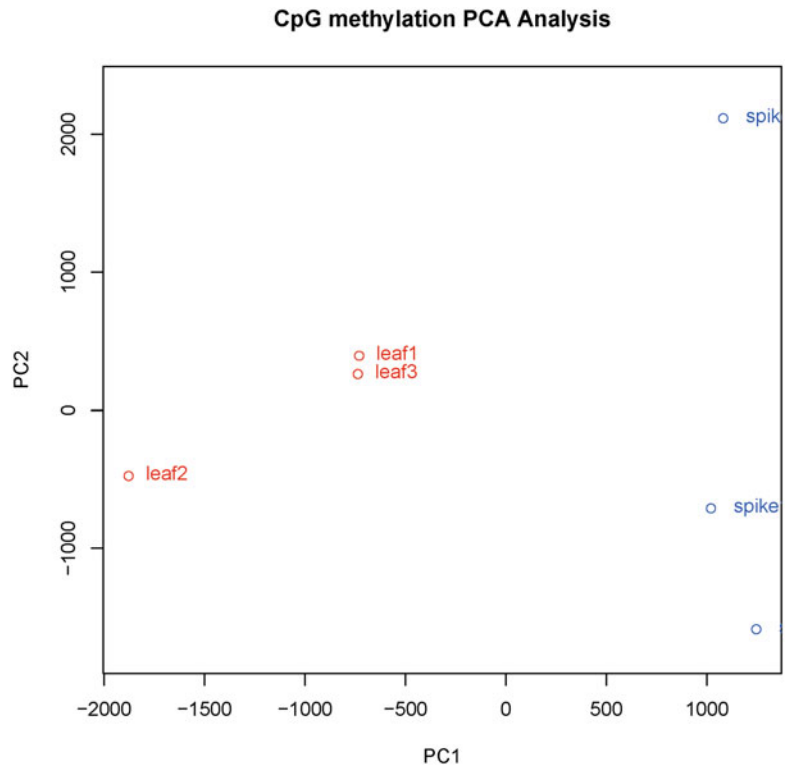

Fig. 6 PCA scatter plot

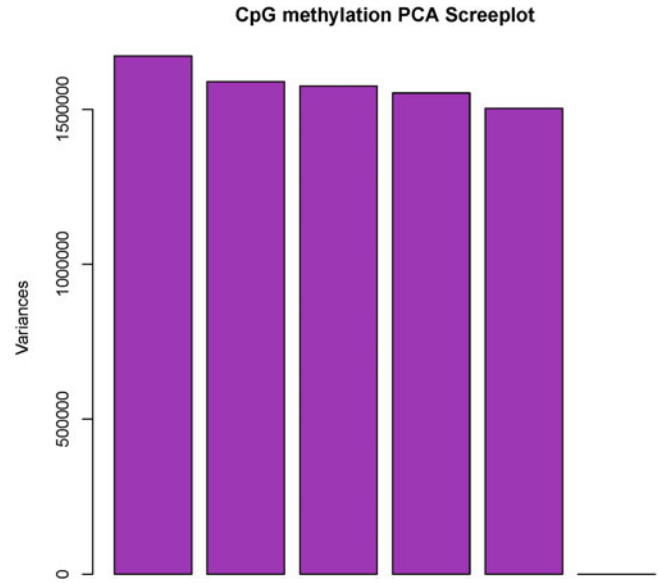

Fig. 7 PCA scree plot

```
3 Bd1  1092 1092      - 0.9456553 0.9211514 -0.37523452
4 Bd1  1176 1176      - 0.4019624 0.9204155  3.80830124
5 Bd1  1560 1560      - 0.9950988 0.9211514 -0.02823264
6 Bd1  1569 1569      - 0.8695866 0.9211514 -0.65231572
```

Then, get.methylDiff function can be used to retrieve differentially methylated bases that satisfy the *Q*-value and percent methylation difference cutoffs. The difference argument defines the cutoff for the methylation difference between test and control (i.e. meth. diff). The qvalue argument defines the cutoff for *Q*-value. Lastly, type argument defines the type of differentially methylated bases returned by the function. Both differential hypomethylated and hypermethylated bases are returned by default (type = "all"), type = "hypo" returns the hypomethylated bases and type = "hyper" returns the hypermethylated bases. Below demonstrate the use of the function on CpG context, please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts.

```
> mDiff25p.cpg = get.methylDiff(mDiff.cpg, difference=25,
qvalue=0.01)
> head(mDiff25p.cpg)
Object of class "methylDiff"
     chr start   end strand      pvalue      qvalue meth.diff
985  Bd1 22420 22420      - 1.775517e-09 2.753255e-06 -44.75043
1330 Bd1 28936 28936      + 1.516198e-09 2.401335e-06 -28.53073
1331 Bd1 28937 28937      - 1.648275e-08 1.897880e-05 -41.54247
2757 Bd1 56058 56058      + 0.000000e+00 0.000000e+00  45.52841
2758 Bd1 56059 56059      - 4.625247e-07 3.304856e-04  45.32258
2886 Bd1 61637 61637      - 7.948650e-08 7.352039e-05 -33.24399

> mDiff25p.hyper.cpg = get.methylDiff(mDiff.cpg,
difference=25, qvalue=0.01, type="hyper")
> head(mDiff25p.hyper.cpg)
Object of class "methylDiff"
     chr start   end strand      pvalue      qvalue meth.diff
2757 Bd1 56058 56058      + 0.000000e+00 0.000000e+00  45.52841
2758 Bd1 56059 56059      - 4.625247e-07 3.304856e-04  45.32258
3100 Bd1 64140 64140      - 2.238908e-09 3.371152e-06  34.88103
3231 Bd1 66299 66299      - 2.419731e-12 8.340650e-09  40.05391
3260 Bd1 66443 66443      + 8.817048e-09 1.108548e-05  27.12750
4647 Bd1 95002 95002      - 3.366775e-06 1.763014e-03  25.42373

> mDiff25p.hypo.cpg = get.methylDiff(mDiff.cpg,
difference=25, qvalue=0.01, type="hypo")
> head(mDiff25p.hypo.cpg)

Object of class "methylDiff"
     chr start   end strand      pvalue      qvalue meth.diff
985  Bd1 22420 22420      - 1.775517e-09 2.753255e-06 -44.75043
```

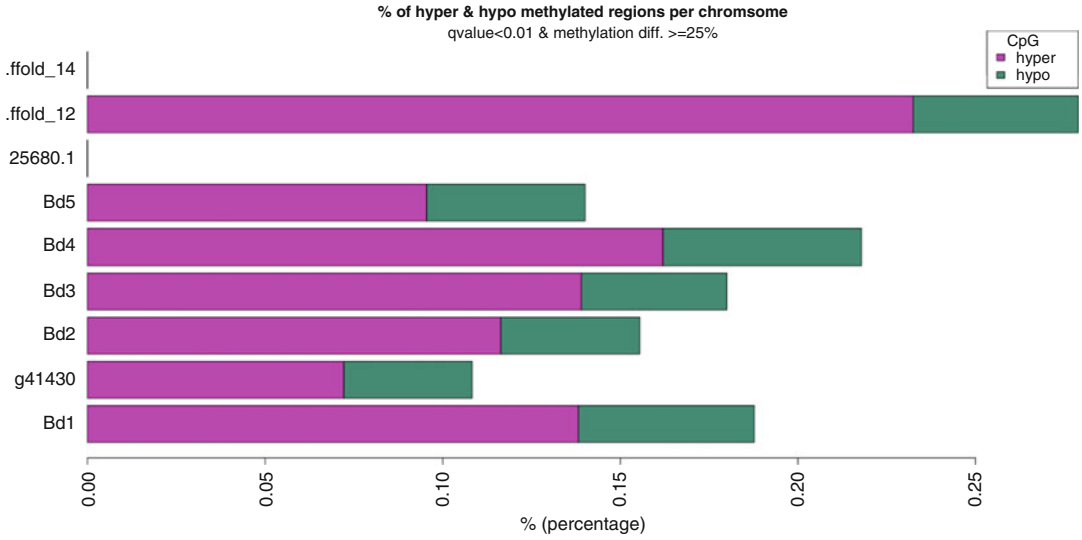

Fig. 8 The proportion of hypomethylated and hypermethylated CpG bases

```
1330 Bd1 28936 28936     + 1.516198e-09 2.401335e-06 -28.53073
1331 Bd1 28937 28937     - 1.648275e-08 1.897880e-05 -41.54247
2886 Bd1 61637 61637     - 7.948650e-08 7.352039e-05 -33.24399
4438 Bd1 91906 91906     - 5.274517e-06 2.562556e-03 -37.15824
4439 Bd1 91908 91908     - 1.420830e-07 1.206177e-04 -36.99495
```

The diffMethPerChr function can be used to print the proportion of differentially methylated CpG bases in each chromosome or produce a horizontal barplots for visualization (*see* Fig. 8).

```
sink("methylKit.DMperChr.cpg.txt")
diffMethPerChr(mDiff.cpg, meth.cutoff=25, qvalue.cutoff=0.01,
plot=FALSE)
sink()

png("methylKit.DMperChr.cpg.png", width=800)
diffMethPerChr(mDiff.cpg, meth.cutoff=25, qvalue.cutoff=0.01,
plot=TRUE)
legend("topright", title="CpG", legend=c("hyper","hypo"),
fill=c("magenta","aquamarine4"))
dev.off()
```

3.6.6 Annotate Differentially Methylated Bases

When given a GRangesList object containing gene annotations such as the promoters, exons, introns and transcription start sites, the annotate.WithGenicParts function can be used in two ways. With a methylDiff object, it annotates the methylation events based on the given gene annotation, and with a GRanges object, it annotates genomic features in the object with annotation.

To prepare the required gene annotation GRangesList object, the read.transcript.features function can be used to read transcript features from a BED file. For genome assemblies available in the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks), the annotation in BED12 format is readily obtainable from its Table Browser. An excerpt of a typical BED file is shown below:

```
chr1    187890    187958    uc031tlm.1    0    -    187890
187890    0    1    68,    0,
chr1    200879    201017    uc031tlo.2    0    +    200879
200879    0    1    138,    0,
chr1    257863    264733    uc057aus.1    0    -    257863
257863    0    2    1162,130,    0,6740,
```

Since *B. distachyon* Bd21 is not available in the UCSC Genome Browser, the gene annotation GRangesList object is build from a TxDB object that has been made available in GitHub. The installation of the R package containing the TxDB object is performed in Subheading 3.6.1. To load the dataset and construct the gene annotation GRangesList object, execute:

```
library("GenomicFeatures")
library(TxDb.Bdistachyon.JGI.Bd3.1.geneexons)
txdb = TxDb.Bdistachyon.JGI.Bd3.1.geneexons

# Build genic GRanges obj
transcripts.GR = transcripts(txdb)
elementMetadata(transcripts.GR) = data.frame(name = element-
Metadata(transcripts.GR)[,2], stringsAsFactors=FALSE)

# Build promoter GRanges obj
promoters.GR = promoters(txdb, upstream=1000, down-
stream=1000)
elementMetadata(promoters.GR) = data.frame(name = elementMe-
tadata(promoters.GR)[,2], stringsAsFactors=FALSE)

# Build TSS GRanges obj
tss.GR = promoters(txdb, upstream=0, downstream=1)
elementMetadata(tss.GR) = data.frame(name = elementMetadata
(tss.GR)[,2], stringsAsFactors=FALSE)# Build exon GRanges obj
exons.GR = unlist(exonsBy(txdb, "tx", use.names=TRUE))
elementMetadata(exons.GR) = data.frame(name = names(exons.
GR), stringsAsFactors=FALSE)
names(exons.GR) = NULL

# Build intron GRanges obj
introns.GR = unlist(intronsByTranscript(txdb, use.name-
s=TRUE))
elementMetadata(introns.GR) = data.frame(name = names(in-
trons.GR), stringsAsFactors=FALSE)
names(introns.GR) = NULL
```

```
# Combine objects
gene.obj = GRangesList("exons" = exons.GR, "introns" = introns.GR, "promoters" = promoters.GR, "TSSes" = tss.GR, "transcripts" = transcripts.GR)

# Save objects for later use
save(gene.obj, file= "gene.obj.RData")
```

Then, annotate.WithGenicParts function is called to calculate the percentage of the differentially methylated CpG bases in promoter, exon, intron and intergenic regions. Please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts.

```
> diffAnn.cpg = annotate.WithGenicParts(mDiff25p.cpg, gene.obj)
> diffAnn.hyper.cpg = annotate.WithGenicParts(mDiff25p.hyper.cpg, gene.obj)
> diffAnn.hypo.cpg = annotate.WithGenicParts(mDiff25p.hypo.cpg, gene.obj)

> diffAnn.cpg
summary of target set annotation with genic parts
28146 rows in target set
_____________
_____________
percentage of target features overlapping with annotation :
  promoter       exon     intron intergenic
  34.22511   30.43416   10.95360   42.53890

percentage of target features overlapping with annotation (with promoter>exon>intron precedence) :
  promoter       exon     intron intergenic
 34.225112  17.657927   5.578057  42.538904

percentage of annotation boundaries with feature overlap :
 promoter      exon    intron
13.673261  2.215517  1.145110

summary of distances to the nearest TSS :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0     717    1596    2933    3252  149200
```

The percentage of the differentially methylated CpG bases in various regions can also be visualized as a pie chart by using the plotTargetAnnotation function (*see* Fig. 9). The precedence argument defines if the hierarchy of annotation features (promoter > exon > intron) is respected when counting the percentage of overlapped bases.

```
png("methylKit.plotTargetGenes.cpg.png")
plotTargetAnnotation(diffAnn.cpg, precedence=TRUE, main="differential CpG methylation annotation")
dev.off()
```

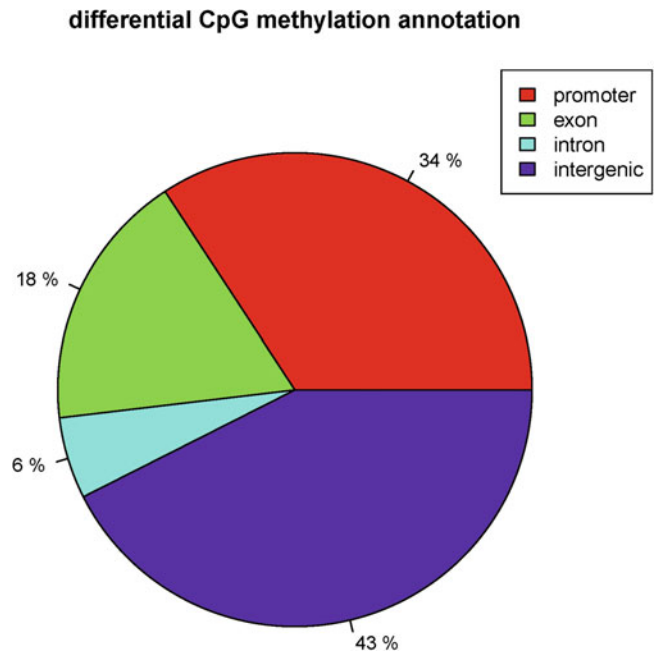

Fig. 9 The proportion of hypomethylated and hypermethylated CpG bases

The getTargetAnnotationStats function returns the number or percentage of target features overlapping with gene annotations. The target features in this case are the differentially methylated CpG bases. The precedence argument defines if the hierarchy of annotation features (promoter > exon > intron) is respected when counting the number or percentage of overlapped features.

```
> getTargetAnnotationStats(diffAnn.cpg, percentage=FALSE,
precedence=TRUE)
  promoter       exon     intron intergenic
      9633       4970       1570      11973
> getTargetAnnotationStats(diffAnn.cpg, percentage=FALSE,
precedence=FALSE)
  promoter       exon     intron intergenic
      9633       8566       3083      11973
```

The getFeatsWithTargetsStats function returns the number or percentage of promoters, exons and introns that overlap with differentially methylated CpG bases.

```
> getFeatsWithTargetsStats(diffAnn.cpg, percentage=FALSE)
promoter     exon   intron
    7243     7034     3029
> getFeatsWithTargetsStats(diffAnn.cpg, percentage=TRUE)
promoter      exon    intron
13.673261  2.215517  1.145110
```

3.6.7 Identify Differentially Methylated Regions

It is also useful to identify differentially methylated regions in the genome where hypermethylated or hypomethylated cytosines occurs consecutively. In such cases, the tileMethylCounts function is used to summarize the methylation information over tiling windows specify size. In the below example, the function tiles the genome with windows of 200 bp in length and 100 bp step-size and calculate the CpG methylation in each tile. Please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts.

```
tiles.cpg = tileMethylCounts(meth.cpg, win.size=200, step.
size=100)
```

This creates a methylBase object with "region" resolution. Then, the same steps of identifying differential hypermethylation and hypomethylation can be applied to the newly created methylBase object.

```
tileDiff.cpg = calculateDiffMeth(tiles.cpg)
tileDiff25p.cpg = get.methylDiff(tileDiff.cpg, difference=25,
qvalue=0.01)
```

Retrieve the number of differentially CpG methylated regions using nrow:

```
> nrow(tileDiff25p.cpg)
[1] 4483
```

Use head function to view part of the content of the object:

```
> head(tileDiff25p.cpg)
methylDiff object with 6 rows
————————————
    chr  start    end strand      pvalue      qvalue meth.diff
273  Bd1  33901  34100          * 0.000000e+00 0.000000e+00 -
25.30420
1959  Bd1  247801  248000        *  3.330669e-16  9.318187e-14
59.24867
2503  Bd1  317201  317400        *  0.000000e+00  0.000000e+00
35.50848
2504  Bd1  317301  317500        *  0.000000e+00  0.000000e+00
29.74244
5069  Bd1  650501  650700        *  5.575427e-06  2.946581e-04
42.95499
7503 Bd1 960201 960400          * 0.000000e+00 0.000000e+00 -
36.26374
————————————
```

Retrieve only the hypermethylated regions:

```
> tileDiff25p.hyper.cpg = get.methylDiff(tileDiff.cpg,
difference=25, qvalue=0.01, type="hyper")
```

```
> head(tileDiff25p.hyper.cpg)
methylDiff object with 6 rows
------------
     chr   start     end strand       pvalue       qvalue meth.
diff
1959 Bd1   247801   248000        * 3.330669e-16 9.318187e-14
59.24867
2503 Bd1   317201   317400        * 0.000000e+00 0.000000e+00
35.50848
2504 Bd1   317301   317500        * 0.000000e+00 0.000000e+00
29.74244
5069 Bd1   650501   650700        * 5.575427e-06 2.946581e-04
42.95499
9371 Bd1  1185901  1186100        * 0.000000e+00 0.000000e+00
33.93221
9372 Bd1  1186001  1186200        * 0.000000e+00 0.000000e+00
44.62039
------------
```

Retrieve only the hypomethylated regions:

```
> tileDiff25p.hypo.cpg = get.methylDiff(tileDiff.cpg,
difference=25, qvalue=0.01, type="hypo")
> head(tileDiff25p.hypo.cpg)
methylDiff object with 6 rows
------------
      chr   start     end strand       pvalue       qvalue meth.
diff
273    Bd1    33901    34100       * 0.000000e+00 0.000000e+00 -
25.30420
7503   Bd1   960201   960400       * 0.000000e+00 0.000000e+00 -
36.26374
14183 Bd1 1814101 1814300       * 1.051951e-05 5.104416e-04 -
30.41958
23619 Bd1 2988301 2988500       * 7.771561e-16 2.089261e-13 -
28.20431
23620 Bd1 2988401 2988600       * 7.771561e-16 2.089261e-13 -
28.20431
24199 Bd1 3063601 3063800       * 3.509517e-05 1.426946e-03 -
28.94737
------------
```

The diffMethPerChr function is then used to print the proportion of differentially methylated regions in each chromosome or produce a horizontal barplots for visualization (*see* Fig. 10).

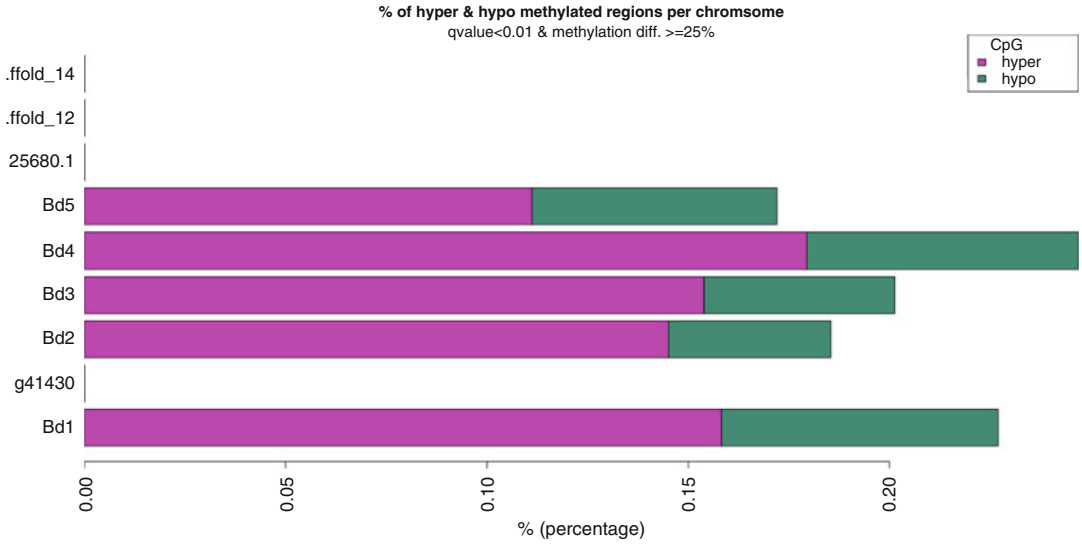

Fig. 10 The proportion of hypomethylated and hypermethylated regions

```
sink("methylKit.DMTperChr.cpg.txt")
diffMethPerChr(tileDiff.cpg, meth.cutoff=25, qvalue.
cutoff=0.01, plot=FALSE)
sink()

png("methylKit.DMTperChr.cpg.png", width=800)
diffMethPerChr(tileDiff.cpg, meth.cutoff=25, qvalue.
cutoff=0.01, plot=TRUE)
legend("topright", title="CpG", legend=c("hyper","hypo"),
fill=c("magenta","aquamarine4"))
dev.off()
```

3.6.8 Annotate Differentially Methylated Regions

As before, the annotate.WithGenicParts function can be used to calculate the percentage of the differentially methylated regions in promoter, exon, intron and intergenic regions. Please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts.

```
> load("gene.obj.RData")

> diffTileAnn.cpg = annotate.WithGenicParts(tileDiff25p.cpg,
gene.obj)
> diffTileAnn.hyper.cpg = annotate.WithGenicParts(tile-
Diff25p.hyper.cpg, gene.obj)
> diffTileAnn.hypo.cpg = annotate.WithGenicParts(tileDiff25p.
hypo.cpg, gene.obj)

> diffTileAnn.cpg
summary of target set annotation with genic parts
4483 rows in target set
_____________
```

```
_____________
percentage of target features overlapping with annotation :
  promoter       exon     intron intergenic
  33.83895   33.23667   26.67856   37.25184

percentage of target features overlapping with annotation
(with promoter>exon>intron precedence) :
  promoter       exon     intron intergenic
  33.83895   20.83426    8.07495   37.25184

percentage of annotation boundaries with feature overlap :
 promoter      exon    intron
2.8448992 0.5889986 0.6241588

summary of distances to the nearest TSS :
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0     700    1636    2668    3197   76910
```

And the plotTargetAnnotation function is used to produce a pie chart as that shown in Fig. 11.

```
png("methylKit.Tile.plotTargetGenes.cpg.png")
plotTargetAnnotation(diffTileAnn.cpg, precedence=TRUE, main-
="regional differential CpG methylation annotation")
dev.off()
```

Lastly, the getTargetAnnotationStats function returns the number or percentage of target features overlapping with gene annotations. The target features in this case are the differentially methylated regions.

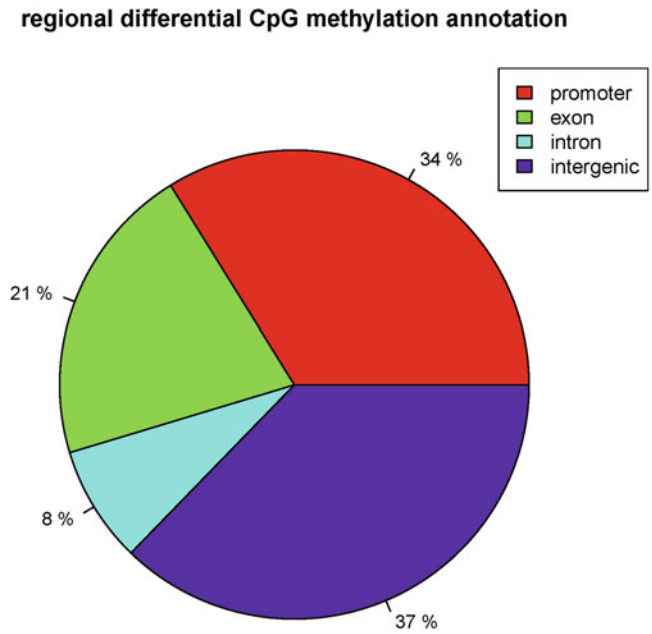

Fig. 11 The proportion of hypomethylated and hypermethylated regions

```
> getTargetAnnotationStats(diffTileAnn.cpg, percentage=FALSE,
precedence=TRUE)
  promoter       exon     intron intergenic
      1517        934        362       1670
> getTargetAnnotationStats(diffTileAnn.cpg, percentage=FALSE,
precedence=FALSE)
  promoter       exon     intron intergenic
      1517       1490       1196       1670
> getFeatsWithTargetsStats(diffTileAnn.cpg, percentage=FALSE)
promoter     exon   intron
    1507     1870     1651
> getFeatsWithTargetsStats(diffTileAnn.cpg, percentage=TRUE)
 promoter      exon    intron
2.8448992 0.5889986 0.6241588
```

3.6.9 Create GRanges Objects of Transposable Elements and Repeats from GFF3

The tandem repeat and transposable element annotations are retrieved from the PGSB/MIPS FTP using wget and decompressed using gunzip:

```
$ wget ftp://ftpmips.helmholtz-muenchen.de/plants/brachypo-
dium/repeats/MIPS_Bd_Transposons_v2.2_16-07-2009.gff3.gz
$ wget ftp://ftpmips.helmholtz-muenchen.de/plants/brachypo-
dium/repeats/MIPS_Bd_tandemRepeats_01-04-2009.gff3.gz
```

```
$ gunzip MIPS_Bd_Transposons_v2.2_16-07-2009.gff3.gz
$ gunzip MIPS_Bd_tandemRepeats_01-04-2009.gff3.gz
```

The MIPS_Bd_tandemRepeats_01–04-2009.gff3 contains a malformed GFF3 header. To fix this bug, please open the file and change:

```
##gff-version    3
```

to:

```
##gff-version 3
```

In the R environment, the import.gff function of the rtracklayer Bioconductor package is used to create GRanges objects from the two GFF files.

First, install rtracklayer by running:

```
> source("http://bioconductor.org/biocLite.R")
> biocLite(c("rtracklayer"))
```

Then, execute:

```
library(BSgenome.Bdistachyon.JGI.Bd3.1)
library(GenomicFeatures)
library(rtracklayer)
```

```
# import files and use unique() to remove any duplicated
entries
tr = import.gff("MIPS_Bd_tandemRepeats_01-04-2009.gff3", gen-
ome="Bdistachyon", format="gff3")
te = import.gff("MIPS_Bd_Transposons_v2.2_16-07-2009.gff3",
genome="Bdistachyon", format="gff3")
tr = unique(tr)
te = unique(te)

# select entries located in the five chromosomes
tr = tr[seqnames(tr) %in% c("Bd1","Bd2","Bd3","Bd4","Bd5")]
te = te[seqnames(te) %in% c("chr01_pseudomolecule","chr02_p-
seudomolecule","chr03_pseudomolecule","chr04_pseudomolecu-
le","chr05_pseudomolecule")]

# drop unwanted levels belonging to scaffolds
tr = dropSeqlevels(tr, levels(seqnames(tr))[6:69])
te = dropSeqlevels(te, levels(seqnames(te))[6:59])

# update TR type
tr$type = paste(tr$type,tr$main_type,sep="_")

# rename to chromosome prefix to "Bd" in TE
te = renameSeqlevels(te, c("Bd1","Bd2","Bd3","Bd4","Bd5"))

# keep the first six columns of the element metadata
elementMetadata(tr) = elementMetadata(tr)[,1:6]
elementMetadata(te) = elementMetadata(te)[,1:6]

# Combine objects and save the GRangesList objects
mips.obj = GRangesList("TR" = tr, "TE" = te)
save(mips.obj, file = "mips.RData")
```

3.6.10 Identify Differentially Methylated Transposable Elements and Repeats

Below, I will demonstrate the use of several methylKit functions to identify differentially methylated tandem repeats (TRs) and transposable elements (TEs) of the CpG context. Please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts. The same work flow can be applied to any region that are stored as GRangesList objects by following the steps listed in Subheading 3.6.9.

The regionCounts function can summarize methylation information (stored as methylRaw, methylRawList or methylBase object) over a set of defined regions. In the below example, the function returns the counts over each TR and TE using a methylRawList object as input.

```
load("mips.RData")
load("methylRawList.RData") # to load objects

# summarize methylation information over TE and TR
TE.cpg = regionCounts(filt.mobj.cpg, mips.obj$TE)
TR.cpg = regionCounts(filt.mobj.cpg, mips.obj$TR)
```

The first two rows of the first sample (i.e. leaf replicate 1) of the TE methylRawList object:

```
> head(TE.cpg[[1]],2)
methylRaw object with 2 rows
_____________
  chr start   end strand coverage numCs numTs
1 Bd1  7511  8542      +      215   198    17
2 Bd1 14817 16722      -     3206  3063   143
_____________
sample.id: leaf1
assembly: Bdistachyon
context: CpG
resolution: region
```

The first two rows of the first sample (i.e. leaf replicate 1) of the TR methylRawList object:

```
> head(TR.cpg[[1]],2)
methylRaw object with 2 rows
_____________
  chr start  end strand coverage numCs numTs
1 Bd1  2503 2751      +       27    24     3
2 Bd1  6159 6434      +       32    32     0
_____________
sample.id: leaf1
assembly: Bdistachyon
context: CpG
resolution: region
```

Then, as before, the unite function is used to produce to a single table:

```
TE_meth.cpg = unite(TE.cpg)
TR_meth.cpg = unite(TR.cpg)
```

Use head function to view part of the content of the two objects:

```
> head(TE_meth.cpg,2)
methylBase object with 2 rows
_____________
   chr start    end strand coverage1 numCs1 numTs1 coverage2
numCs2 numTs2
1 Bd1  7511  8542      +       215    198     17       193    188
5
2 Bd1 14817 16722      -      3206   3063    143      3077   2929
148
```

```
  coverage3 numCs3 numTs3 coverage4 numCs4 numTs4 coverage5
numCs5 numTs5
1       196    186     10       210    197     13       180    168
12
2      3677   3480    197      3170   2985    185      3130   2983
147
  coverage6 numCs6 numTs6
1       175    166       9
2      2279   2156     123
_____________
sample.ids: leaf1 leaf2 leaf3 spike1 spike2 spike3
destranded FALSE
assembly: Bdistachyon
context: CpG
treament: 0 0 0 1 1 1
resolution: region

> head(TR_meth.cpg,2)
methylBase object with 2 rows
_____________
   chr  start   end  strand  coverage1  numCs1  numTs1  coverage2
numCs2 numTs2
1 Bd1  2503 2751      +        27     24      3        13     13
0
2 Bd1  6159 6434      +        32     32      0        38     36
2
   coverage3 numCs3 numTs3 coverage4 numCs4 numTs4 coverage5
numCs5 numTs5
1        21     21      0        18     18      0        27     27
0
2       165    154     11        68     64      4        76     76
0
  coverage6 numCs6 numTs6
1        11     11       0
2        39     35       4
_____________
sample.ids: leaf1 leaf2 leaf3 spike1 spike2 spike3
destranded FALSE
assembly: Bdistachyon
context: CpG
treament: 0 0 0 1 1 1
resolution: region
```

The differential methylation statistics is calculated with the calculateDiffMeth function and get.methylDiff function is used to retrieve the differentially methylated TEs and TRs that satisfy the Q-value and percent methylation difference cutoffs:

```
TEdiff.cpg = calculateDiffMeth(TE_meth.cpg)
TRdiff.cpg = calculateDiffMeth(TR_meth.cpg)
```

```
TEdiff25p.cpg = get.methylDiff(TEdiff.cpg, difference=25,
qvalue=0.01)
TRdiff25p.cpg = get.methylDiff(TRdiff.cpg, difference=25,
qvalue=0.01)
```

Use head function to view part of the content of the two objects:

```
> head(TEdiff25p.cpg,2)
methylDiff object with 2 rows
_____________
     chr    start      end strand       pvalue       qvalue meth.
diff
1037 Bd1   6093414   6093464       - 2.828521e-08 4.463105e-07
39.62500
2131 Bd1 11007451 11007561      + 5.154128e-06 5.744991e-05 -
32.63403
_____________
sample.ids: leaf1 leaf2 leaf3 spike1 spike2 spike3
destranded FALSE
assembly: Bdistachyon
context: CpG
treament: 0 0 0 1 1 1
resolution: region

> head(TRdiff25p.cpg,2)
methylDiff object with 2 rows
_____________
    chr   start    end strand      pvalue      qvalue meth.diff
259  Bd1  2855651  2855716         +  7.172041e-13  1.363628e-10
57.90323
357  Bd1  3806587  3806612         +  1.981257e-06  1.220958e-04
28.58586
_____________
sample.ids: leaf1 leaf2 leaf3 spike1 spike2 spike3
destranded FALSE
assembly: Bdistachyon
context: CpG
treament: 0 0 0 1 1 1
resolution: region
```

Retrieve the number of differentially methylated TEs and TRs nrow:

```
> nrow(TEdiff25p.cpg)
[1] 132

> nrow(TRdiff25p.cpg)
[1] 88
```

Retrieve the number of differentially hypermethylated and hypomethylated TEs and TRs:

```
> nrow(get.methylDiff(TEdiff.cpg, difference=25, qvalue=0.01,
type="hyper"))
[1] 85

> nrow(get.methylDiff(TEdiff.cpg, difference=25, qvalue=0.01,
type="hypo"))
[1] 47

> nrow(get.methylDiff(TRdiff.cpg, difference=25, qvalue=0.01,
type="hyper"))
[1] 65

> nrow(get.methylDiff(TRdiff.cpg, difference=25, qvalue=0.01,
type="hypo"))
[1] 23
```

3.6.11 Plot the Proportion of Differentially Methylated Transposable Elements and Repeats

In this subsection, I will use the ggplot2 package to produce barplots to show the percentage of differentially methylated transposable elements and tandem repeats, and also boxplots to shows the distribution of methylation difference of the three cytosine contexts.

In the R environment, load the required packages with library and create a data.frame object named TR containing all the entries of tandem repeats and the corresponding differential methylation information.

```
library(GenomicRanges)
library(data.table)
library(ggplot2)

TR = merge(as.data.frame(mips.obj$TR)[,c(1:3,7)], data.frame
(TRdiff.cpg)[,c(1:3,6,7)], by.x = c("seqnames","star-
t","end"), by.y = c("chr","start","end"), all.x = T)
TR = merge(TR, data.frame(TRdiff.chg)[,c(1:3,6,7)], by.x = c
("seqnames","start","end"), by.y = c("chr","start","end"),
all.x = T)
TR = merge(TR, data.frame(TRdiff.chh)[,c(1:3,6,7)], by.x = c
("seqnames","start","end"), by.y = c("chr","start","end"),
all.x = T)
```

Update the column name and types of tandem repeat.

```
names(TR)[5:10] = c("q.cpg","d.cpg","q.chg","d.chg","q.
chh","d.chh")
TR$type = as.factor(as.character(TR$type))
levels(TR$type) = c("TR_micro","TR_mini","TR_sat","TR_vntr")
```

Create additional columns to specify if a tandem repeat is differentially methylated (q-value <0.05).

```
TR$s.cpg = "nDM"
TR[TR$q.cpg < 0.05 & TR$d.cpg < 0 & !is.na(TR$q.cpg),]$s.cpg =
"Hypo"
TR[TR$q.cpg < 0.05 & TR$d.cpg > 0 & !is.na(TR$q.cpg),]$s.cpg =
"Hyper"
TR$s.chg = "nDM"
TR[TR$q.chg < 0.05 & TR$d.chg < 0 & !is.na(TR$q.chg),]$s.chg =
"Hypo"
TR[TR$q.chg < 0.05 & TR$d.chg > 0 & !is.na(TR$q.chg),]$s.chg =
"Hyper"
TR$s.chh = "nDM"
TR[TR$q.chh < 0.05 & TR$d.chh < 0 & !is.na(TR$q.chh),]$s.chh =
"Hypo"
TR[TR$q.chh < 0.05 & TR$d.chh > 0 & !is.na(TR$q.chh),]$s.chh =
"Hyper"
```

Create a data.frame object named percTR to store the fraction of hypermethylated and hypomethylated tandem repeats.

```
percTR = rbind(data.frame(context = "CpG", meth = c("Hy-
per","Hypo"), freq = prop.table(table(TR$s.cpg))[1:2], row.
names = NULL),
data.frame(context = "CHG", meth = c("Hyper","Hypo"), freq =
prop.table(table(TR$s.chg))[1:2], row.names = NULL),
data.frame(context = "CHH", meth = c("Hyper","Hypo"), freq =
prop.table(table(TR$s.chh))[1:2], row.names = NULL))
```

Call the ggplot function to create barplots (*see* Fig. 12).

```
png("percTR.png", height = 2000, width = 2500, res = 300)
ggplot(percTR, aes(meth, freq*100, fill = meth)) +
geom_bar(stat = "identity", color = "black", size = 1, width =
0.8) +
guides(fill = guide_legend(override.aes = list(size = 0.5)))
+
facet_wrap(~ context) + theme_light() + theme(legend.posi-
tion="right",
axis.title.x = element_blank(),
axis.title.y = element_text(face = "bold", size = 16),
axis.text.x = element_text(size = 14), axis.text.y = ele-
ment_text(size = 14),
strip.text.x = element_text(face = "bold", size = 16, color =
"black"),
legend.title = element_blank(), panel.grid.minor = element_-
blank(),
plot.title = element_text(face = "bold", size = 18)) +
labs(title = "Proportion of differentially methylated tandem
repeats", y = "%")
dev.off()
```

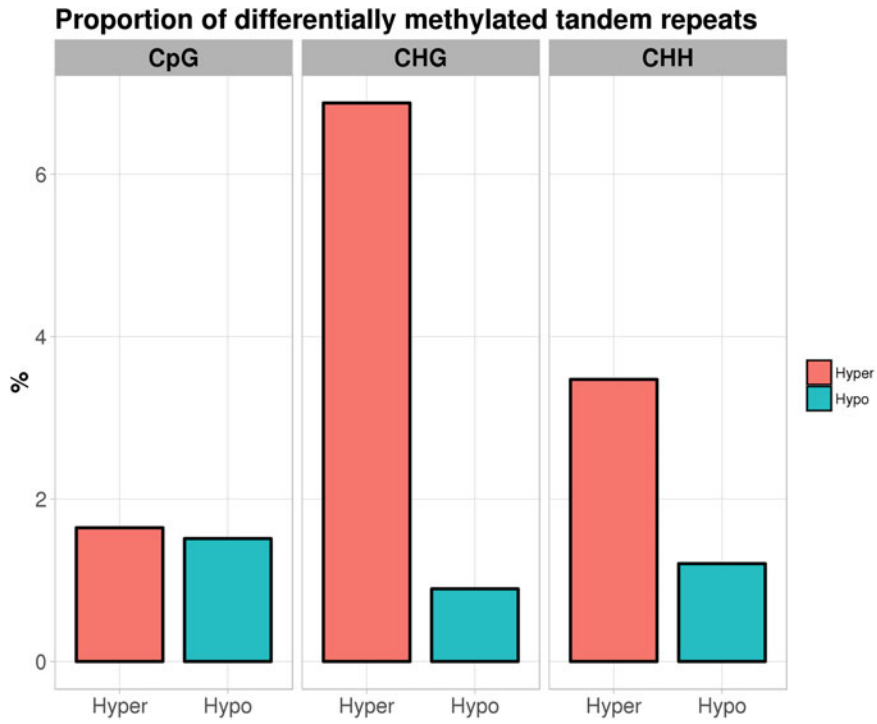

Fig. 12 The proportion of hypomethylated and hypermethylated tandem repeats

Similarly, executing the below commands to create the barplots that showed the proportion of differentially methylated transposable elements (*see* Fig. 13).

```
TE = merge(as.data.frame(mips.obj$TE)[,c(1:3,7)], data.frame
(TEdiff.cpg)[,c(1:3,6,7)], by.x = c("seqnames","star-
t","end"), by.y = c("chr","start","end"), all.x = T)
TE = merge(TE, data.frame(TEdiff.chg)[,c(1:3,6,7)], by.x = c
("seqnames","start","end"), by.y = c("chr","start","end"),
all.x = T)
TE = merge(TE, data.frame(TEdiff.chh)[,c(1:3,6,7)], by.x = c
("seqnames","start","end"), by.y = c("chr","start","end"),
all.x = T)

names(TE)[5:10] = c("q.cpg","d.cpg","q.chg","d.chg","q.
chh","d.chh")
TE$type = as.factor(as.character(TE$type))

TE$s.cpg = "nDM"
TE[TE$q.cpg < 0.05 & TE$d.cpg < 0 & !is.na(TE$q.cpg),]$s.cpg =
"Hypo"
TE[TE$q.cpg < 0.05 & TE$d.cpg > 0 & !is.na(TE$q.cpg),]$s.cpg =
"Hyper"
TE$s.chg = "nDM"
TE[TE$q.chg < 0.05 & TE$d.chg < 0 & !is.na(TE$q.chg),]$s.chg =
```

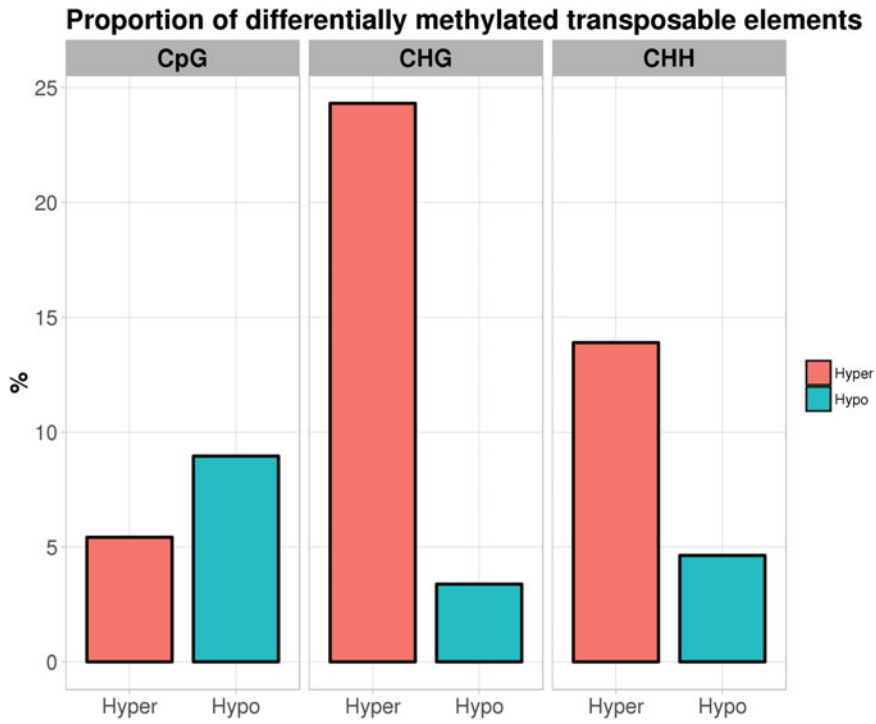

Fig. 13 The proportion of hypomethylated and hypermethylated transposable elements

```
"Hypo"
TE[TE$q.chg < 0.05 & TE$d.chg > 0 & !is.na(TE$q.chg),]$s.chg =
"Hyper"
TE$s.chh = "nDM"
TE[TE$q.chh < 0.05 & TE$d.chh < 0 & !is.na(TE$q.chh),]$s.chh =
"Hypo"
TE[TE$q.chh < 0.05 & TE$d.chh > 0 & !is.na(TE$q.chh),]$s.chh =
"Hyper"

percTE = rbind(data.frame(context = "CpG", meth = c("Hy-
per","Hypo"), freq = prop.table(table(TE$s.cpg))[1:2], row.
names = NULL),
data.frame(context = "CHG", meth = c("Hyper","Hypo"), freq =
prop.table(table(TE$s.chg))[1:2], row.names = NULL),
data.frame(context = "CHH", meth = c("Hyper","Hypo"), freq =
prop.table(table(TE$s.chh))[1:2], row.names = NULL))

png("percTE.png", height = 2000, width = 2500, res = 300)
ggplot(percTE, aes(meth, freq*100, fill = meth)) +
geom_bar(stat = "identity", color = "black", size = 1, width =
0.8) +
guides(fill = guide_legend(override.aes = list(size = 0.5)))
+
facet_wrap(~ context) + theme_light() + theme(legend.posi-
tion="right",
```

```
axis.title.x = element_blank(),
axis.title.y = element_text(face = "bold", size = 16),
axis.text.x = element_text(size = 14), axis.text.y = ele-
ment_text(size = 14),
strip.text.x = element_text(face = "bold", size = 16, color =
"black"),
legend.title = element_blank(), panel.grid.minor = element_-
blank(),
plot.title = element_text(face = "bold", size = 18)) +
labs(title = "Proportion of differentially methylated trans-
posable elements", y = "%")
dev.off()
```

A data.frame object named dataTR is created to store the differentially methylated tandem repeats and the difference in methylation.

```
dataTR = rbind(data.frame(type = TR[TR$s.cpg == "Hyper",]
$type, context = "CpG", diff = "Hyper", meth = TR[TR$s.cpg
== "Hyper",]$d.cpg),
data.frame(type = TR[TR$s.cpg == "Hypo",]$type, context =
"CpG", diff = "Hypo", meth = TR[TR$s.cpg == "Hypo",]$d.cpg),
data.frame(type = TR[TR$s.chg == "Hyper",]$type, context =
"CHG", diff = "Hyper", meth = TR[TR$s.chg == "Hyper",]$d.
chg),
data.frame(type = TR[TR$s.chg == "Hypo",]$type, context =
"CHG", diff = "Hypo", meth = TR[TR$s.chg == "Hypo",]$d.chg),
data.frame(type = TR[TR$s.chh == "Hyper",]$type, context =
"CHH", diff = "Hyper", meth = TR[TR$s.chh == "Hyper",]$d.
chh),
data.frame(type = TR[TR$s.chh == "Hypo",]$type, context =
"CHH", diff = "Hypo", meth = TR[TR$s.chh == "Hypo",]$d.chh))
```

Call the ggplot function to create boxplots (*see* Fig. 14).

```
png("TR_meth.png", height = 3000, width = 2500, res = 300)
ggplot(dataTR, aes(type, meth, fill = diff)) +
geom_boxplot(color = "black", size = 0.5, width = 0.8, out-
lier.size = 0.1, outlier.color = "blue") +
guides(fill = guide_legend(override.aes = list(size = 0.5)))
+
facet_wrap(~ context, ncol = 1) + theme_light() + theme
(legend.position="right",
axis.title.x = element_blank(),
axis.title.y = element_text(face = "bold", size = 16),
axis.text.x = element_text(size = 14), axis.text.y = ele-
ment_text(size = 14),
strip.text.x = element_text(face = "bold", size = 16, color =
"black"),
legend.title = element_blank(), panel.grid.minor = element_-
blank(),
```

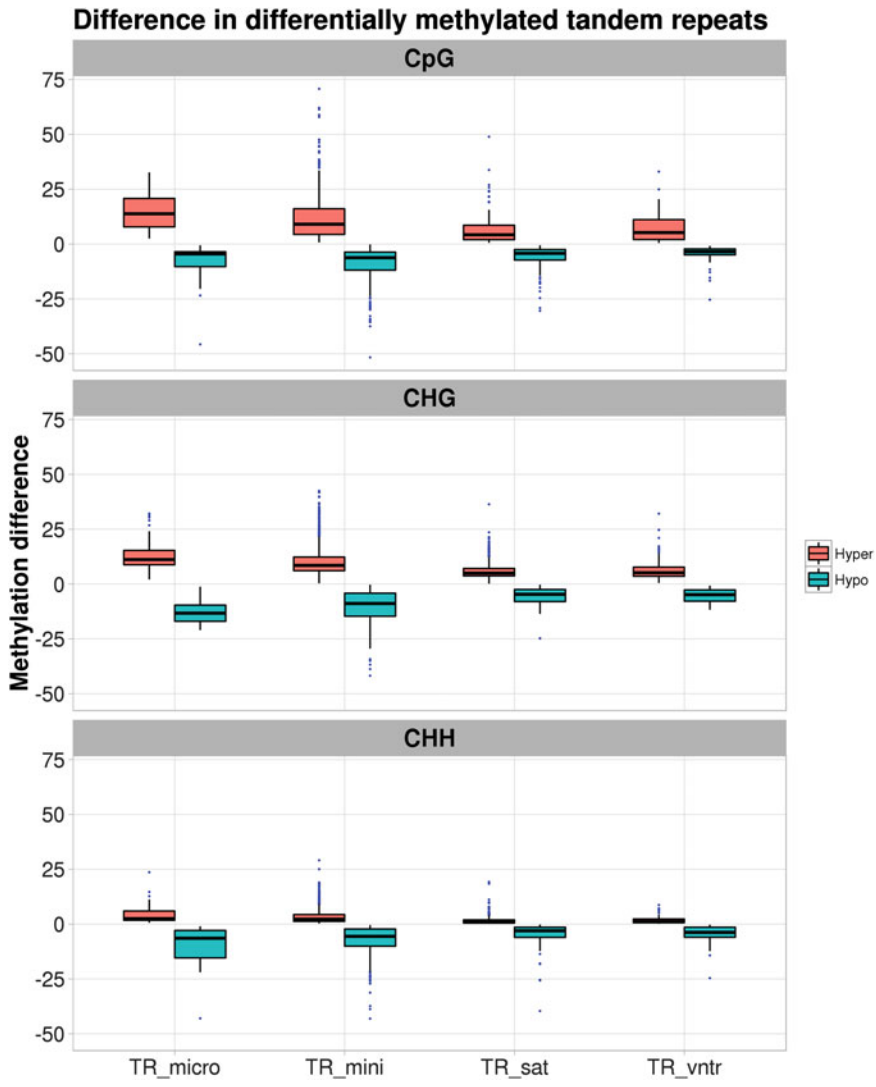

Fig. 14 The distribution of methylation changes in the differentially methylated tandem repeats

```
plot.title = element_text(face = "bold", size = 18)) +
labs(title = "Difference in differentially methylated tandem
repeats", y = "Methylation difference")
dev.off()
```

Similarly, executing the below commands to create the boxplots that showed the distribution of the difference in methylation in the differentially methylated transposable elements (*see* Fig. 15).

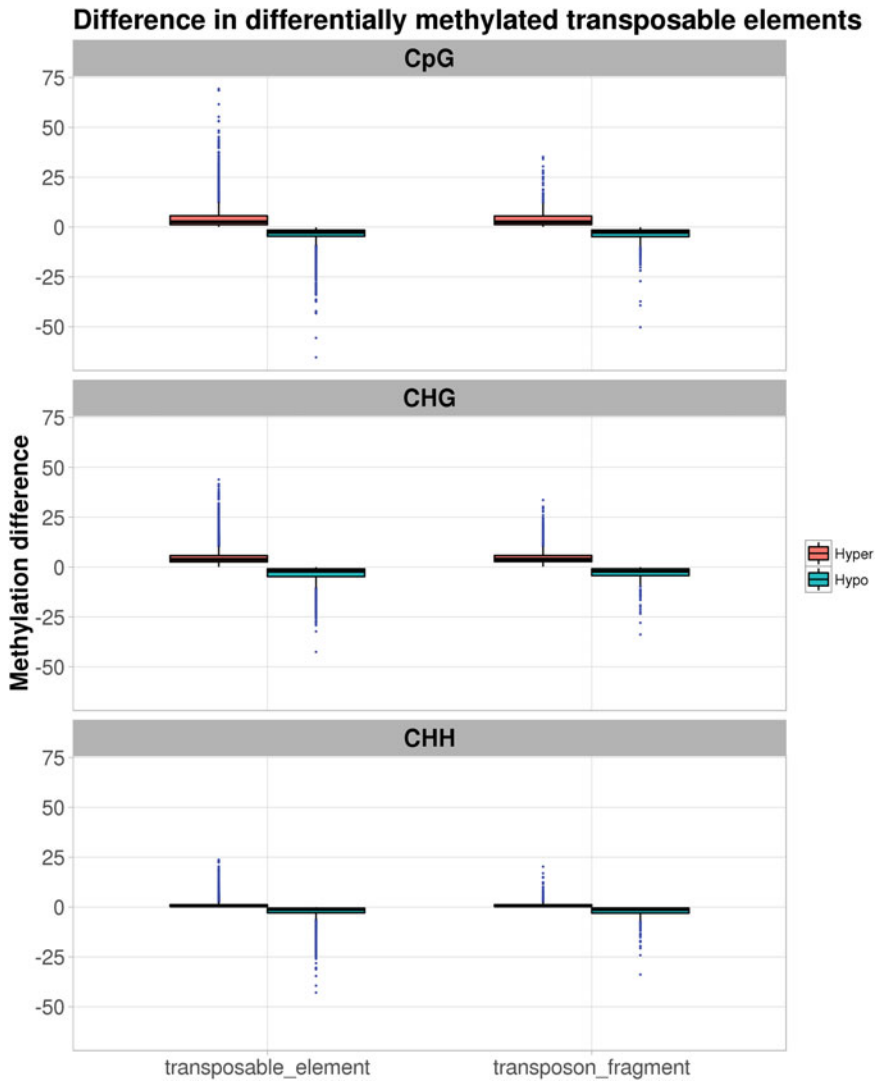

Fig. 15 The distribution of methylation changes in the differentially methylated transposable elements

```
dataTE = rbind(data.frame(type = TE[TE$s.cpg == "Hyper",]
$type, context = "CpG", diff = "Hyper", meth = TE[TE$s.cpg
== "Hyper",]$d.cpg),
data.frame(type = TE[TE$s.cpg == "Hypo",]$type, context =
"CpG", diff = "Hypo", meth = TE[TE$s.cpg == "Hypo",]$d.cpg),
data.frame(type = TE[TE$s.chg == "Hyper",]$type, context =
"CHG", diff = "Hyper", meth = TE[TE$s.chg == "Hyper",]$d.
chg),
data.frame(type = TE[TE$s.chg == "Hypo",]$type, context =
```

```
"CHG", diff = "Hypo", meth = TE[TE$s.chg == "Hypo",]$d.chg),
data.frame(type = TE[TE$s.chh == "Hyper",]$type, context =
"CHH", diff = "Hyper", meth = TE[TE$s.chh == "Hyper",]$d.
chh),
prop.table(table(TE$s.chh))[1:2], row.names = NULL))
png("percTE.png", height = 2000, width = 2500, res = 300)
ggplot(dataTE, aes(type, meth, fill = diff)) +
geom_boxplot(color = "black", size = 0.5, width = 0.8, out-
lier.size = 0.1, outlier.color = "blue") +
scale_y_continuous(breaks = c(seq(-50,75,25)), labels = seq(-
50,75,25)) +
guides(fill = guide_legend(override.aes = list(size = 0.5)))
+
facet_wrap(~ context, ncol = 1) + theme_light() + theme
(legend.position="right",
axis.title.x = element_blank(),
axis.title.y = element_text(face = "bold", size = 16),
axis.text.x = element_text(size = 14), axis.text.y = ele-
ment_text(size = 14),
strip.text.x = element_text(face = "bold", size = 16, color =
"black"),
legend.title = element_blank(), panel.grid.minor = element_-
blank(),
plot.title = element_text(face = "bold", size = 18)) +
labs(title = "Difference in differentially methylated trans-
posable elements", y = "Methylation difference")
dev.off()
```

3.6.12 Other Functions

The methylBase object stores the number of reads containing cytosine and thymine at each base in each sample. The percMethylation function calculates and returns a matrix object containing the percent methylation at each base in the methylBase object. For example:

```
perc.meth.cpg = percMethylation(meth.cpg)
perc.meth.chg = percMethylation(meth.chg)
perc.meth.chh = percMethylation(meth.chh)
```

Retrieve the dimension of the matrices with dim function:

```
> dim(perc.meth.cpg)
[1] 15783683        6

> dim(perc.meth.chg)
[1] 12799098        6

> dim(perc.meth.chh)
[1] 31497439        6
```

Use head function to view part of the content of the objects:

```
> head(perc.meth.cpg,2)
      leaf1     leaf2     leaf3    spike1    spike2    spike3
1  94.82759  89.39394  95.34884  93.33333  98.14815  93.15068
2  94.44444 100.00000 100.00000 100.00000 100.00000  94.11765

> head(perc.meth.chg,2)
     leaf1   leaf2    leaf3   spike1    spike2    spike3
1 74.54545 76.5625 80.72289 91.07143  96.15385  91.04478
2 96.96970 95.0000 96.96970 90.00000 100.00000 100.00000

> head(perc.meth.chh,2)
     leaf1    leaf2    leaf3   spike1   spike2 spike3
1 1.724138 1.538462 4.705882 1.694915 1.886792      0
2 0.000000 0.000000 0.000000 0.000000 0.000000      0
```

The percent methylation at each TE and TR can also be obtained using the percMethylation function. Below shows the commands to retrieve CpG methylation:

```
perc.TE_meth.cpg = percMethylation(TE_meth.cpg)
perc.TR_meth.cpg = percMethylation(TR_meth.cpg)
```

Retrieve the dimension of the matrices with dim function:

```
> dim(perc.TE_meth.cpg)
[1] 73229     6

> dim(perc.TR_meth.cpg)
[1] 27144     6
```

Use head function to view part of the content of the objects:

```
> head(perc.TE_meth.cpg,2)
     leaf1    leaf2    leaf3   spike1   spike2   spike3
1 92.09302 97.40933 94.89796 93.80952 93.33333 94.85714
2 95.53961 95.19012 94.64237 94.16404 95.30351 94.60290

> head(perc.TR_meth.cpg,2)
      leaf1     leaf2     leaf3    spike1 spike2    spike3
1  88.88889 100.00000 100.00000 100.00000    100 100.00000
2 100.00000  94.73684  93.33333  94.11765    100  89.74359
```

The methylation values of the six samples can be useful for alternative analysis or visualization purposes. In the next section, I will show how the methylation values can be presented as line graphs over gene body, as well as heatmap representation. For now, the base-resolution matrices are save into a RData file:

```
save(perc.meth.cpg, perc.meth.chg, perc.meth.chh, file=
"perc.meth.RData")
```

The as function can coercing a methylRaw, methylBase and methylDiff object to a GRanges object, for example to coerce a methylBase object, run:

```
> GR = as(meth.cpg,"GRanges")
> head(GR,3)
GRanges object with 3 ranges and 18 metadata columns:
       seqnames         ranges strand | coverage1     numCs1
numTs1 coverage2
          <Rle>     <IRanges>  <Rle> | <integer> <numeric>
<numeric> <integer>
  [1]      Bd1 [ 260,  260]      - |        58        55         3
66
  [2]      Bd1 [ 354,  354]      - |        18        17         1
22
  [3]      Bd1 [1092, 1092]      - |        15        14         1
11
       numCs2    numTs2 coverage3    numCs3    numTs3 coverage4
numCs4
        <numeric> <numeric> <integer> <numeric> <numeric>
<integer> <numeric>
  [1]        59         7        86        82         4        60
56
  [2]        22         0        39        39         0        13
13
  [3]        10         1        15        14         1        18
16
       numTs4 coverage5    numCs5    numTs5 coverage6    numCs6
numTs6
        <numeric> <integer> <numeric> <numeric> <integer>
<numeric> <numeric>
  [1]         4        54        53         1        73        68
5
  [2]         0        13        13         0        17        16
1
  [3]         2        23        21         2        11        11
0
  ———
   seqinfo: 9 sequences from an unspecified genome; no
seqlengths
```

And to coerce a methylDiff object:

```
> GR = as(mDiff25p.cpg,"GRanges")
> head(GR,3)
GRanges object with 3 ranges and 2 metadata columns:
      seqnames         ranges strand |                qvalue
meth.diff
```

```
        <Rle>      <IRanges>  <Rle> |           <numeric>
<numeric>
  [1]      Bd1 [22420, 22420]      - | 2.75325521400437e-06  -
44.750430292599
  [2]      Bd1 [28936, 28936]      + | 2.40133456106072e-06 -
28.5307285307285
  [3]      Bd1 [28937, 28937]      - | 1.89787996335248e-05 -
41.5424739195231
  -----
   seqinfo: 7 sequences from an unspecified genome; no
seqlengths
```

3.7 Analyze BS-seq Data with R/ Bioconductor Package EnrichedHeatmap

The EnrichedHeatmap package produces heatmaps to represent the enrichment of genomic signals at regions of interest, for example at transcription start sites, CpG island, gene body, etc. [6]. It consists of a two-step process: (1) calculation of the association between genomic signals and target regions by normalizing to a matrix, and (2) creation of the heatmap to represent the values stored in the matrix. Below I will use the matrices produced from the first step of the process and the ggplot2 package to create line graphs to show CpG methylation profiles. I will also show the use of the EnrichedHeatmap function to create heatmaps. Please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts.

3.7.1 Installation of R Packages

In R environment, install the latest version of reshape2 from CRAN by running:

```
install.packages("reshape2")
```

Install EnrichedHeatmap and circlize from GitHub using the install_github function of devtools:

```
devtools::install_github("jokergoo/EnrichedHeatmap")
devtools::install_github("jokergoo/circlize")
```

3.7.2 Create GRanges Objects

Execute the below commands to load two previously saved R objects and create a GRanges object for each sample to store the base-resolution CpG methylation data.

```
library(GenomicRanges)
load("perc.meth.RData") # perc.meth.cpg, perc.meth.chg, perc.
meth.chh
load("methylBase.RData") # meth.cpg, meth.chg, meth.chh

cpgGR = granges(as(meth.cpg, "GRanges"))
leaf1.cpgGR = leaf2.cpgGR = leaf3.cpgGR = spike1.cpgGR =
spike2.cpgGR = spike3.cpgGR = cpgGR
```

```
elementMetadata(leaf1.cpgGR)$meth = perc.meth.cpg[,1]
elementMetadata(leaf2.cpgGR)$meth = perc.meth.cpg[,2]
elementMetadata(leaf3.cpgGR)$meth = perc.meth.cpg[,3]
elementMetadata(spike1.cpgGR)$meth = perc.meth.cpg[,4]
elementMetadata(spike2.cpgGR)$meth = perc.meth.cpg[,5]
elementMetadata(spike3.cpgGR)$meth = perc.meth.cpg[,6]
```

3.7.3 Normalize Associations Between Absolute Methylation Levels and Region Around Transcription Start Sites (TSS ± 2.5 kb) into Matrix

Next, the normalizeToMatrix function is called to calculate the association between the levels of CpG methylation and region around transcription start sites (TSS ± 2.5 kb) by normalizing to a matrix. The amount of extension from the TSS is specified with the extend parameter, and the value_column parameter specifies which column in the GRanges object stores the methylation data. The default window size (w) is calculated as max(extend)/50, and can be changed by using a different w value. In each window, how the mean methylation value is calculated is controlled by the mean_mode parameter, which can be one of the following: absolute, weighted and w0. Smoothing may be enabled by using smooth = TRUE. to improve the visualization of final output.

```
library("EnrichedHeatmap")
load("gene.obj.RData") # gene.obj
extend = 2500

TSS_leaf1.cpg = normalizeToMatrix(leaf1.cpgGR, gene.obj
$TSSes, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA)
TSS_leaf2.cpg = normalizeToMatrix(leaf2.cpgGR, gene.obj
$TSSes, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA)
TSS_leaf3.cpg = normalizeToMatrix(leaf3.cpgGR, gene.obj
$TSSes, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA)
TSS_spike1.cpg = normalizeToMatrix(spike1.cpgGR, gene.obj
$TSSes, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA)
TSS_spike2.cpg = normalizeToMatrix(spike2.cpgGR, gene.obj
$TSSes, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA)
TSS_spike3.cpg = normalizeToMatrix(spike3.cpgGR, gene.obj
$TSSes, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA)
```

3.7.4 Calculate Column-Wise Mean and Plot Line Graphs Around TSS ± 2.5 kb

The colMeans function is applied to each normalizedMatrix object to calculate column-wise mean, i.e. mean methylation of each window. Then, a data.frame object (dfTSS) is created to store the column-wise mean of all samples.

```
library(reshape2)
library(ggplot2)
```

```
mean_TSS_leaf1.cpg = colMeans(TSS_leaf1.cpg, na.rm = TRUE)
mean_TSS_leaf2.cpg = colMeans(TSS_leaf2.cpg, na.rm = TRUE)
mean_TSS_leaf3.cpg = colMeans(TSS_leaf3.cpg, na.rm = TRUE)
mean_TSS_spike1.cpg = colMeans(TSS_spike1.cpg, na.rm = TRUE)
mean_TSS_spike2.cpg = colMeans(TSS_spike2.cpg, na.rm = TRUE)
mean_TSS_spike3.cpg = colMeans(TSS_spike3.cpg, na.rm = TRUE)

dfTSS = data.frame(Sample = c("Leaf1"," Leaf2"," Leaf3","Spi-
ke1","Spike2","Spike3"))
dfTSS[,names(mean_TSS_leaf1.cpg)] <- rbind(mean_TSS_leaf1.
cpg, mean_TSS_leaf2.cpg, mean_TSS_leaf3.cpg, mean_TSS_spike1.
cpg, mean_TSS_spike2.cpg, mean_TSS_spike3.cpg)
dfTSS = melt(dfTSS, id = c("Sample"))
```

Call the ggplot function to create line graphs showing the change in mean methylation values around transcription start site (*see* Fig. 16).

```
png("promoter_CpG_methylation.png", width = 1000)
ggplot(dfTSS, aes(variable, value, group = Sample)) +
geom_line(aes(color = Sample), size = 0.6) + theme_light() +
scale_x_discrete(breaks = c("u1","d50"), labels=c("-
2500","+2500")) +
theme(legend.position="right",
axis.title.x = element_text(face = "bold", size = 16),
axis.title.y = element_text(face = "bold", size = 16),
axis.text.x = element_text(size = 14), axis.text.y = ele-
ment_text(size = 14),
panel.grid.minor = element_blank(),
plot.title = element_text(face = "bold", size = 18)) +
labs(title = "Average CpG methylation at promoter region", x =
```

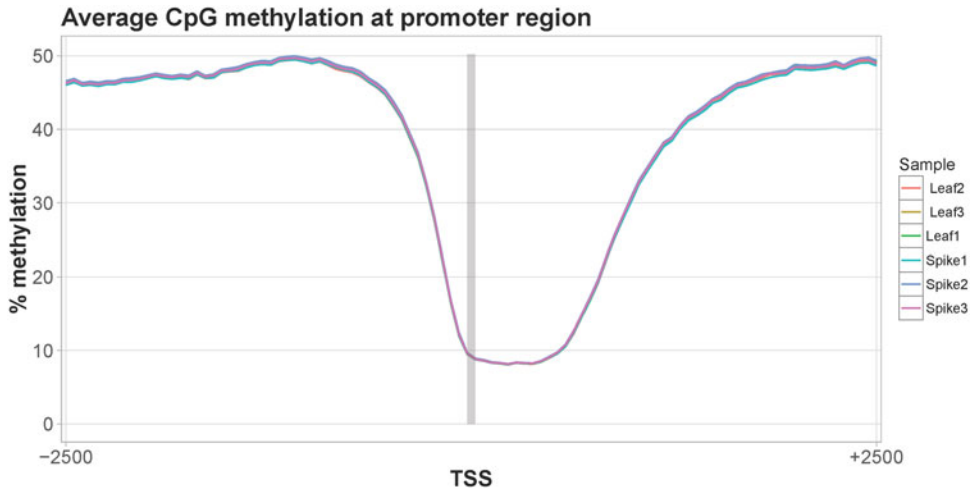

Fig. 16 Averaged CpG methylation levels surrounding the transcription start site. The target region is shaded with *gray box*

```
"TSS", y = "% methylation") +
annotate("rect", xmin = 50, xmax = 51, ymin = 0, ymax = max
(dfTSS$value)+0.25, alpha = .2)
dev.off()
```

3.7.5 Normalize Associations Between Absolute Methylation Levels and Transcribed Region (Transcript ± 2.5 kb) into Matrix

Similarly, the normalizeToMatrix function is called to calculate the association between the levels of CpG methylation and region around the transcripts (transcripts ± 2.5 kb) by normalizing to a matrix. The target_ratio parameter controls the width of the target regions, i.e. the transcribed regions in this example, relative to the full heatmap. Here, I set the target_ratio to 0.4 so that ratio of the width of the upstream and downstream 2.5 kb regions in the heatmap will be 0.3 each.

```
Tx_leaf1.cpg = normalizeToMatrix(leaf1.cpgGR, gene.obj$tran-
scripts, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA, target_ratio = 0.4)
Tx_leaf2.cpg = normalizeToMatrix(leaf2.cpgGR, gene.obj$tran-
scripts, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA, target_ratio = 0.4)
Tx_leaf3.cpg = normalizeToMatrix(leaf3.cpgGR, gene.obj$tran-
scripts, value_column = "meth", mean_mode = "absolute", extend
= extend, empty_value = NA, target_ratio = 0.4)
Tx_spike1.cpg = normalizeToMatrix(spike1.cpgGR, gene.obj
$transcripts, value_column = "meth", mean_mode = "absolute",
extend = extend, empty_value = NA, target_ratio = 0.4)
Tx_spike2.cpg = normalizeToMatrix(spike2.cpgGR, gene.obj
$transcripts, value_column = "meth", mean_mode = "absolute",
extend = extend, empty_value = NA, target_ratio = 0.4)
Tx_spike3.cpg = normalizeToMatrix(spike3.cpgGR, gene.obj
$transcripts, value_column = "meth", mean_mode = "absolute",
extend = extend, empty_value = NA, target_ratio = 0.4)
```

3.7.6 Calculate Column-Wise Mean and Plot Line Graphs around Transcript ± 2.5 kb

Execute the following commands to calculate the column-wise mean and create the data.frame object (dfTx) containing these values:

```
mean_Tx_leaf1.cpg = colMeans(Tx_leaf1.cpg, na.rm = TRUE)
mean_Tx_leaf2.cpg = colMeans(Tx_leaf2.cpg, na.rm = TRUE)
mean_Tx_leaf3.cpg = colMeans(Tx_leaf3.cpg, na.rm = TRUE)
mean_Tx_spike1.cpg = colMeans(Tx_spike1.cpg, na.rm = TRUE)
mean_Tx_spike2.cpg = colMeans(Tx_spike2.cpg, na.rm = TRUE)
mean_Tx_spike3.cpg = colMeans(Tx_spike3.cpg, na.rm = TRUE)

dfTx = data.frame(Sample = c("Leaf1"," Leaf2"," Leaf3","Spi-
ke1","Spike2","Spike3"))
dfTx[,names(mean_Tx_leaf1.cpg)] <- rbind(mean_Tx_leaf1.cpg,
mean_Tx_leaf2.cpg, mean_Tx_leaf3.cpg, mean_Tx_spike1.cpg,
mean_Tx_spike2.cpg, mean_Tx_spike3.cpg)
dfTx = melt(dfTx, id = c("Sample"))
```

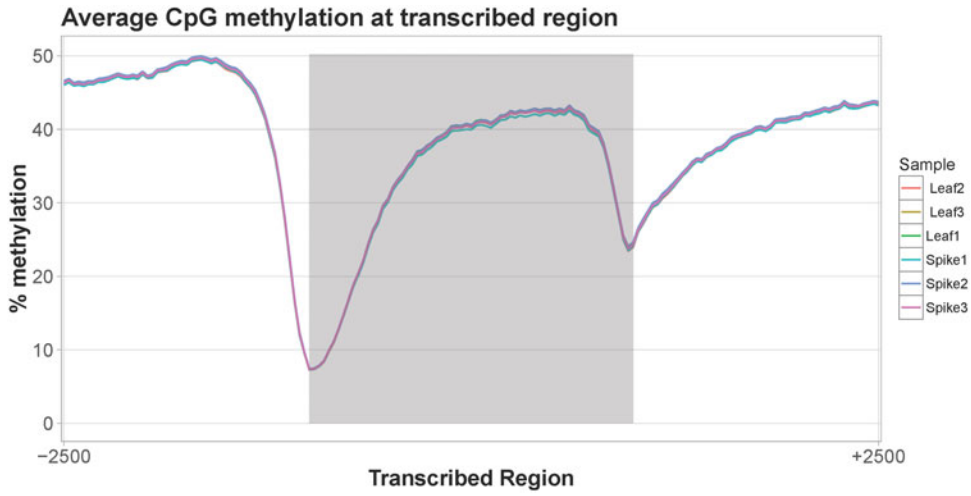

Fig. 17 Averaged CpG methylation levels along the transcribed region. The target region is shaded with *gray box*

Call the ggplot function to create line graphs showing the change in mean methylation values around gene body (*see* Fig. 17).

```
png("tx_CpG_methylation.png", width = 1000)
ggplot(dfTx, aes(variable, value, group = Sample)) +
geom_line(aes(color = Sample), size = 0.6) + theme_light() +
scale_x_discrete(breaks = c("u1","d50"), labels=c("-
2500","+2500")) + theme(legend.position="right",
axis.title.x = element_text(face = "bold", size = 16),
axis.title.y = element_text(face = "bold", size = 16),
axis.text.x = element_text(size = 14), axis.text.y = ele-
ment_text(size = 14),
panel.grid.minor = element_blank(),
plot.title = element_text(face = "bold", size = 18)) +
labs(title = "Average CpG methylation at transcribed region ",
x = "Transcribed Region", y = "% methylation") +
annotate("rect", xmin = 51, xmax = 117, ymin = 0, ymax = max
(dfTx$value)+0.25, alpha = .2)
dev.off()
```

3.7.7 Create Enrichment Heatmap

First, the necessary libraries are loaded. The divisive hierarchical clustering algorithm (diana) from the cluster package will be used to perform clustering of rows, i.e. regions around transcription start sites. The colorRamp2 function of the circlize package is used to built the a color mapping function by specifying vectors of breaks values and corresponding colors.

```
library(cluster) # diana
library(circlize) # colorRamp2
methcol = colorRamp2(c(0, 0.5, 1), c("blue", "white", "red"))
```

The heatmap corresponding to the CpG methylation profile around transcription start site is generate by calling the EnrichedHeatmap function. To demonstrate the plotting of the methylation profile of chromosome Bd5, the subset of the normalizedMatrix object is used as input. I use the table function to build a contingency table of the counts of transcription start sites in each chromosome.

```
> data.frame(table(seqnames(gene.obj$TSSes)))
                                      Var1  Freq
1                                      Bd1 15192
2                                      Bd2 11974
3                                      Bd3 11734
4                                      Bd4  8960
5                                      Bd5  5096
6                               scaffold_12     6
7                               scaffold_14     1
8                              scaffold_135     1
9                              scaffold_180     1
10 Bd1_centromere_containing_Bradi1g41430     7
```

The entries corresponding to Bd5 are between row number 47,861–52,956.

```
> gene.obj$TSSes[47861:52956,]
GRanges object with 5096 ranges and 1 metadata column:
         seqnames              ranges strand   |           name
           <Rle>           <IRanges>  <Rle>   |    <character>
     [1]      Bd5     [230482, 230482]      +   | Bradi5g00220.3
     [2]      Bd5     [240216, 240216]      +   | Bradi5g00230.2
     [3]      Bd5     [273630, 273630]      +   | Bradi5g00274.1
     [4]      Bd5     [277065, 277065]      +   | Bradi5g00297.3
     [5]      Bd5     [277065, 277065]      +   | Bradi5g00297.4
     ...      ...                  ...    ... ...            ...
  [5092]      Bd5 [28615776, 28615776]      -   | Bradi5g27720.1
  [5093]      Bd5 [28618117, 28618117]      -   | Bradi5g27720.5
  [5094]      Bd5 [28618132, 28618132]      -   | Bradi5g27720.6
  [5095]      Bd5 [28618695, 28618695]      -   | Bradi5g27720.3
  [5096]      Bd5 [28618117, 28618117]      -   | Bradi5g27720.4
  -----
  seqinfo: 10 sequences from an unspecified genome
```

Hence, to subset the normalizedMatrix object to containing only data from Bd5 can be done with the following command:

```
> TSS_leaf1.cpg[47861:52956,]
Normalize leaf1.cpgGR to gene.obj$TSSes:
  Upstream 2500 bp (50 windows)
  Downstream 2500 bp (50 windows)
  Not include target regions
  5096 signal regions
```

When creating the heatmap, clustering of the rows is by default disabled, and may be enabled by specifying cluster_rows = TRUE in the EnrichedHeatmap function to perform hierarchical clustering with hclust. Here, I demonstrate how one can bypass the use of the default clustering algorithm by generating a dendrogram object and feed into the EnrichedHeatmap function. Empty values in the matrix is assigned a value of 0.5 to avoid errors when performing distance matrix calculation. Executing the following commands to produce a dendrogram object:

```
mat = TSS_leaf1.cpg[47861:52956,]
mat[is.na(mat)] = 0.5
dend = as.dendrogram(diana(mat))
```

Then, the EnrichedHeatmap is called. The dendrogram object created above is assigned to the cluster_rows parameter and set the row_dend_reorder parameter to FALSE to disable reordering of the dendrogram. The axis_name parameter is defined to reflect the upstream and downstream extension from the transcription start site. Several EnrichedHeatmap objects can be joined to display side-by-side. Execute the following to built a EnrichedHeatmapList object:

```
HT_TSS.cpg = EnrichedHeatmap(TSS_leaf1.cpg[47861:52956,],
cluster_rows = dend, row_dend_reorder = FALSE, col = methcol,
name = "Leaf1 CpG meth", column_title = "Leaf1 CpG", axis_name
= c("-2.5Kb", "TSS", "2.5Kb")) +
EnrichedHeatmap(TSS_leaf2.cpg[47861:52956,], col = methcol,
name = "Leaf2 CpG meth", column_title = "Leaf2 CpG", axis_name
= c("-2.5Kb", "TSS", "2.5Kb")) +
EnrichedHeatmap(TSS_leaf3.cpg[47861:52956,], col = methcol,
name = "Leaf3 CpG meth", column_title = "Leaf3 CpG", axis_name
= c("-2.5Kb", "TSS", "2.5Kb")) +
EnrichedHeatmap(TSS_spike1.cpg[47861:52956,], col = methcol,
name = "Spike1 CpG meth", column_title = "Spike1 CpG",
axis_name = c("-2.5Kb", "TSS", "2.5Kb")) +
EnrichedHeatmap(TSS_spike2.cpg[47861:52956,], col = methcol,
name = "Spike2 CpG meth", column_title = "Spike2 CpG",
axis_name = c("-2.5Kb", "TSS", "2.5Kb")) +
EnrichedHeatmap(TSS_spike3.cpg[47861:52956,], col = methcol,
name = "Spike3 CpG meth", column_title = "Spike3 CpG",
axis_name = c("-2.5Kb", "TSS", "2.5Kb"))
```

Call the draw function to create heatmaps (*see* Fig. 18).

```
png("promoter_CpG_methylation_ht.png", height = 3000, width =
4800, res = 300)
draw(HT_TSS.cpg, gap = unit(1, "cm"))
dev.off()
```

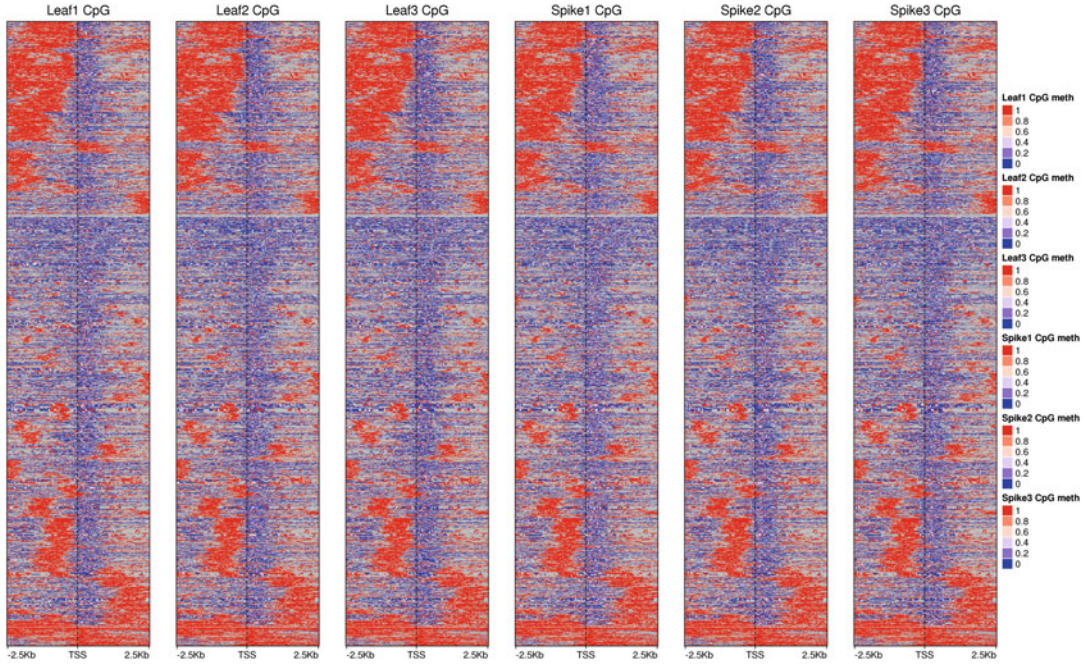

Fig. 18 CpG methylation profiles around all transcription start sites in chromosome Bd5

Similarly, clustering was performed before creating the heatmap to show the methylation profile over transcribed regions, as follow:

```
mat = Tx_leaf1.cpg[47861:52956,]
mat[is.na(mat)] = 0.5
dend = as.dendrogram(diana(mat))
```

The axis_name parameter is defined differently here to reflect the upstream and downstream extension from the transcription start site and termination site respectively. Execute the following to built a EnrichedHeatmapList object:

```
HT_Tx.cpg = EnrichedHeatmap(Tx_leaf1.cpg[47861:52956,], cluster_rows = dend, row_dend_reorder = FALSE, col = methcol, name = "Leaf1 CpG meth", column_title = "Leaf1 CpG", axis_name = c("-2.5Kb", "Start", "End", "2.5Kb")) +
EnrichedHeatmap(Tx_leaf2.cpg[47861:52956,], col = methcol, name = "Leaf2 CpG meth", column_title = "Leaf2 CpG", axis_name = c("-2.5Kb", "Start", "End", "2.5Kb")) +
EnrichedHeatmap(Tx_leaf3.cpg[47861:52956,], col = methcol, name = "Leaf3 CpG meth", column_title = "Leaf3 CpG", axis_name = c("-2.5Kb", "Start", "End", "2.5Kb")) +
EnrichedHeatmap(Tx_spike1.cpg[47861:52956,], col = methcol, name = "Spike1 CpG meth", column_title = "Spike1 CpG",
```

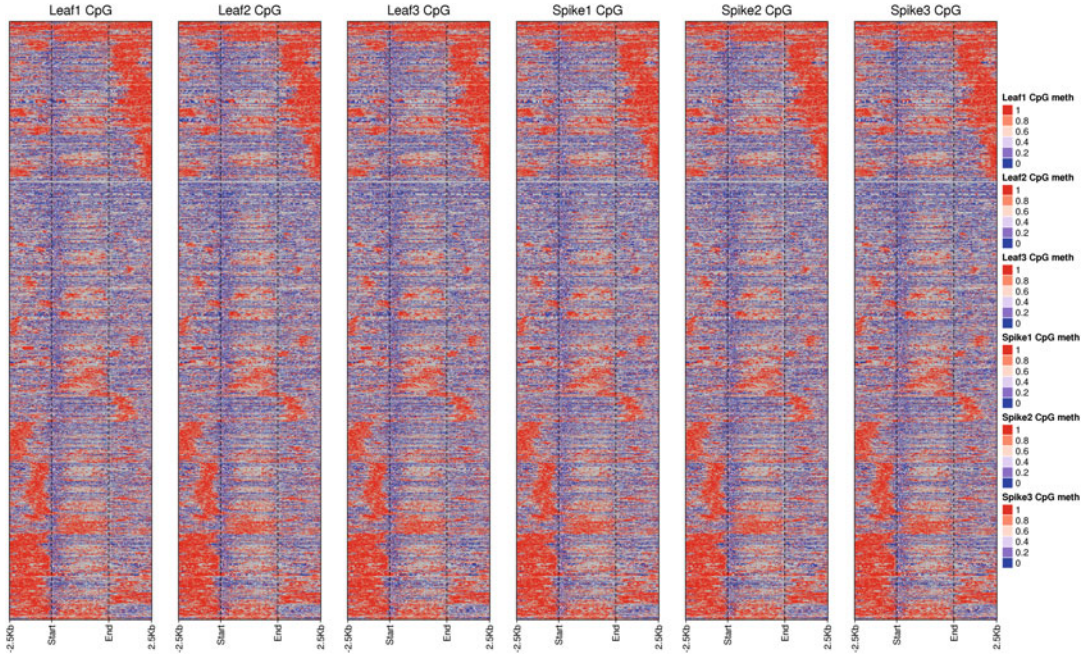

Fig. 19 CpG methylation profiles around all transcribed regions in chromosome Bd5

```
axis_name = c("-2.5Kb", "Start", "End", "2.5Kb")) +
EnrichedHeatmap(Tx_spike2.cpg[47861:52956,], col = methcol,
name = "Spike2 CpG meth", column_title = "Spike2 CpG",
axis_name = c("-2.5Kb", "Start", "End", "2.5Kb")) +
EnrichedHeatmap(Tx_spike3.cpg[47861:52956,], col = methcol,
name = "Spike3 CpG meth", column_title = "Spike3 CpG",
axis_name = c("-2.5Kb", "Start", "End", "2.5Kb"))
```

Call the draw function to create heatmaps (*see* Fig. 19).

```
png("tx_CpG_methylation_ht.png", height = 3000, width = 4800,
res = 300)
draw(HT_Tx.cpg, gap = unit(1, "cm"))
dev.off()
```

3.8 Analyze BS-seq Data with R/Bioconductor Package methylPipe

Below, I will present an alternative method of analyzing BS-seq data using methylPipe [7]. The package works with BS-seq data from organisms that are mapped to UCSC assemblies and use UCSC annotations. That means, some of its functions are not fully compatible with this dataset. I have created a fork of methylPipe and released it in the GitHub repository. Please make sure all the necessary packages are installed before proceeding.

3.8.1 Installation of R Packages

Install the required (and dependent) packages from GitHub using the install_github function of devtools:

```
devtools::install_github("ycl6/methylPipe")
devtools::install_github("ycl6/BSgenome.Bdistachyon.JGI.Bd3.1")
devtools::install_github("ycl6/TxDb.Bdistachyon.JGI.Bd3.1.geneexons")
```

3.8.2 Prepare the Methylation Call Files for methylPipe

The methylPipe package take the input tab-delimited file containing the following columns: (1) chromosome name, (2) genomic position, (3) strand, (4) methylation context (CG, CHG or CHH), (5) number of reads with C calls, and (6) number of reads with T calls. By executing the brlow command, the *.methcount file is converted into the format required by methylPipe, and each context from each sample is placed in individual folder.

```
mkdir methylPipe methylPipe/CpG methylPipe/CHG methylPipe/CHH
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438 SRR629439
do
     mkdir methylPipe/CpG/${id} methylPipe/CHG/${id} methylPipe/CHH/${id}
   awk -F $'\t' 'BEGIN { OFS=FS } { C = sprintf("%.0f",$5*$6); T = $6-C; if(($4 == "CpG" || $4 == "CpGx") && C > 0) print $1,$2+1,$3,"CG",C,T }' ${id}.methcount > methylPipe/CpG/${id}/${id}.txt
   awk -F $'\t' 'BEGIN { OFS=FS } { C = sprintf("%.0f",$5*$6); T = $6-C; if(($4 == "CHH" || $4 == "CHHx") && C > 0) print $1,$2+1,$3,"CHH",C,T }' ${id}.methcount > methylPipe/CHH/${id}/${id}.txt
   awk -F $'\t' 'BEGIN { OFS=FS } { C = sprintf("%.0f",$5*$6); T = $6-C; if(($4 == "CCG" || $4 == "CCGx" || $4 == "CXG" || $4 == "CXGx") && C > 0) print $1,$2+1,$3,"CHG",C,T }' ${id}.methcount > methylPipe/CHG/${id}/${id}.txt
done
```

3.8.3 Prepare the Uncovered (uncov) Files for methylPipe

methylPipe also requires information regarding genomic regions not covered in the BS-seq data. This information can be produced by running genomeCoverageBed from BedTools with the MR alignment generated previously by MethPipe (Subheading 3.2) to identify regions that were not sequenced. The -bga option reports the depth in BedGraph format, including regions with zero coverage. Execute the following command to produce the "uncov" files:

```
for id in SRR628921 SRR629088 SRR629207 SRR629437 SRR629438 SRR629439
do
```

```
    genomeCoverageBed -i ${id}.dremove -g chromInfo.txt -bga |
awk -F $'\t' 'BEGIN { OFS=FS } { if($4 == 0) print $1,$2+1,$3,
$4 }' > methylPipe/${id}.uncov.bed
done
```

The chromInfo.txt file used in the above command contains the chromosomal length information of *B. distachyon* Bd21:

```
$ cat chromInfo.txt
Bd1     75071545
Bd2     59130575
Bd3     59640145
Bd4     48594894
Bd5     28630136
scaffold_12     23566
scaffold_14     20560
scaffold_135    3881
scaffold_180    1933
Bd1_centromere_containing_Bradi1g41430  46184
EU325680.1      135199
```

3.8.4 Reading of Methylation Call Files

The BSprepare function from methylPipe is called to processing the methylation data and the output is stored as a TABIX-compressed and indexed file. Here the addchr option is set as FALSE so that the word "chr" is not added to the chromosome name (i.e. Bd1 does not become chrBd1). In R environment, execute:

```
library(methylPipe)

tabix = "path_to_tabix_folder" # e.g. /pkg/samtools_1.2+/
htslib-1.2.1
for(id in c("SRR628921", "SRR629088", "SRR629207",
"SRR629437", "SRR629438", "SRR629439")) {
    cpgLoc = paste("methylPipe/CpG/",id, sep="")
    BSprepare(cpgLoc, cpgLoc, tabix=tabix, addchr=FALSE)
    chgLoc = paste("methylPipe/CHG/",id, sep="")
    BSprepare(chgLoc, chgLoc, tabix=tabix, addchr=FALSE)
    chhLoc = paste("methylPipe/CHH/",id, sep="")
    BSprepare(chhLoc, chhLoc, tabix=tabix, addchr=FALSE)
}
```

After completion, three files are produced in each folder, for example the three files in methylPipe/CpG/SRR628921 are:

1. SRR628921_tabix_out.txt.gz
2. SRR628921_tabix_out.txt.gz.tbi
3. SRR628921_tabix.txt.

3.8.5 Genome-Wide DNA Methylation Profiling

In the remaining subsections, the CpG context is used to demonstrate the analysis of BS-seq data with methylPipe. Please refer to Supplementary Files 1 and 2 for corresponding analysis with the CHG and CHH contexts.

To load the required libraries and define variables:

```
library(data.table) #fread
library(GenomicFeatures)
library(BSgenome.Bdistachyon.JGI.Bd3.1) # Bdistachyon
library(TxDb.Bdistachyon.JGI.Bd3.1.geneexons)

leaf1 = "SRR628921"; leaf2 = "SRR629088"; leaf3 = "SRR629207"
spike1 = "SRR629437"; spike2 = "SRR629438"; spike3 =
"SRR629439"

uncovPath = "methylPipe"
mathPath = "methylPipe/CpG"
```

Load the files containing uncovered regions and convert to GRanges objects:

```
leaf1_uncov <- data.frame(fread(paste(uncovPath,"/",leaf1,".
uncov.bed", sep = ""), header = F))
leaf2_uncov <- data.frame(fread(paste(uncovPath,"/",leaf2,".
uncov.bed", sep = ""), header = F))
leaf3_uncov <- data.frame(fread(paste(uncovPath,"/",leaf3,".
uncov.bed", sep = ""), header = F))
spike1_uncov <- data.frame(fread(paste(uncovPath,"/",
spike1,".uncov.bed", sep = ""), header = F))
spike2_uncov <- data.frame(fread(paste(uncovPath,"/",
spike2,".uncov.bed", sep = ""), header = F))
spike3_uncov <- data.frame(fread(paste(uncovPath,"/",
spike3,".uncov.bed", sep = ""), header = F))
```

Assign column names:

```
names(leaf1_uncov) = names(leaf2_uncov) = names(leaf3_uncov)
= names(spike1_uncov) = names(spike2_uncov) = names(spi-
ke3_uncov) = c("chr","start","end","score")
```

Convert data.frame objects to GRanges objects:

```
leaf1_uncov = makeGRangesFromDataFrame(leaf1_uncov)
leaf2_uncov = makeGRangesFromDataFrame(leaf2_uncov)
leaf3_uncov = makeGRangesFromDataFrame(leaf3_uncov)
spike1_uncov = makeGRangesFromDataFrame(spike1_uncov)
spike2_uncov = makeGRangesFromDataFrame(spike2_uncov)
spike3_uncov = makeGRangesFromDataFrame(spike3_uncov)
```

The BSdata object is created to store DNA methylation data and the BSdataSet object stores a collection of BSdata objects. To create BSdata objects:

```
leaf1.db <- BSdata(file=paste(mathPath,"/",leaf1,"/",
leaf1,"_tabix_out.txt.gz", sep = ""), uncov=leaf1_uncov,
org=Bdistachyon)
leaf2.db <- BSdata(file=paste(mathPath,"/",leaf2,"/",
leaf2,"_tabix_out.txt.gz", sep = ""), uncov=leaf2_uncov,
org=Bdistachyon)
leaf3.db <- BSdata(file=paste(mathPath,"/",leaf3,"/",
leaf3,"_tabix_out.txt.gz", sep = ""), uncov=leaf3_uncov,
org=Bdistachyon)
spike1.db <- BSdata(file=paste(mathPath,"/",spike1,"/",spi-
ke1,"_tabix_out.txt.gz", sep = ""), uncov=spike1_uncov,
org=Bdistachyon)
spike2.db <- BSdata(file=paste(mathPath,"/",spike2,"/",spi-
ke2,"_tabix_out.txt.gz", sep = ""), uncov=spike2_uncov,
org=Bdistachyon)
spike3.db <- BSdata(file=paste(mathPath,"/",spike3,"/",spi-
ke3,"_tabix_out.txt.gz", sep = ""), uncov=spike3_uncov,
org=Bdistachyon)
```

Create BSdataSet object:

```
Bdistachyon.cpg <- BSdataSet(org=Bdistachyon, group=c(rep
("C",3),rep("E",3)), leaf1=leaf1.db, leaf2=leaf2.db,
leaf3=leaf3.db, spike1=spike1.db, spike2=spike2.db, spi-
ke3=spike3.db)
```

The mCsmoothing function can plot methylation profile over a given genomic region. The smoothing can be performed by calculating the "sum" or "mean" of methylation level for each regions. In the below example, the mean methylation profile is plotted over each of the five chromosomes (Scorefun parameter), and each chromosome is divided into 5000 bins (Nbins parameter). The graphical outputs were organized and illustrated in Fig. 20.

```
Bdchr = data.frame(seqnames = seqnames(Bdistachyon), seq-
lengths = seqlengths(Bdistachyon), stringsAsFactors = FALSE)
Bdchr = Bdchr[1:5,] # Keep 5 major chromosomes

for (i in c(1:5)) {
        chr = Bdchr$seqnames[i]
        len = Bdchr$seqlengths[i]
            pdf(paste("mCsmoothing-",chr,".cpg.pdf", sep=""),
width=10, height=10, pointsize=12)
           mCsmoothing(leaf1.db, GRanges(chr,IRanges(1,len)),
Scorefun='mean', Nbins=5000, Context="CG", plot=TRUE)
           mCsmoothing(leaf2.db, GRanges(chr,IRanges(1,len)),
Scorefun='mean', Nbins=5000, Context="CG", plot=TRUE)
           mCsmoothing(leaf3.db, GRanges(chr,IRanges(1,len)),
```

Fig. 20 CpG methylation profile along all chromosomes

```
Scorefun='mean', Nbins=5000, Context="CG", plot=TRUE)
          mCsmoothing(spike1.db, GRanges(chr,IRanges(1,len)),
Scorefun='mean', Nbins=5000, Context="CG", plot=TRUE)
          mCsmoothing(spike2.db, GRanges(chr,IRanges(1,len)),
Scorefun='mean', Nbins=5000, Context="CG", plot=TRUE)
          mCsmoothing(spike3.db, GRanges(chr,IRanges(1,len)),
Scorefun='mean', Nbins=5000, Context="CG", plot=TRUE)
        dev.off()
}
```

The chromosome is divided into 5000 bins and smoothing is performed to calculate the average methylation level within each bin.

3.8.6 Descriptive Statistics About the Methylation Data

The methstats function can calculate basic statistics such as mean, median and quantile distribution of methylation of a BSdataSet object. It also computes the pairwise Pearson correlation

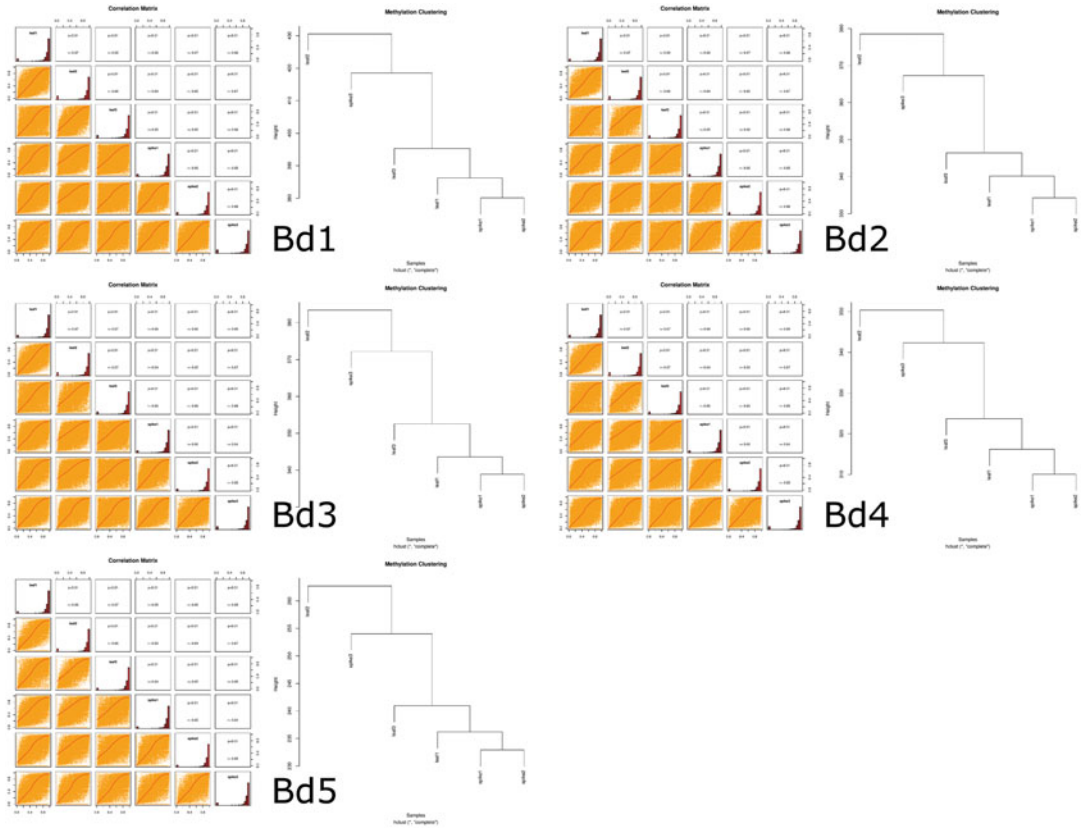

Fig. 21 CpG methylation similarity between samples assessed with pairwise Pearson correlation coefficients and hierarchical clustering

coefficients of the methylation profiles between samples, and performs hierarchical clustering. Execute the below commands to save the scatter plot and dendrogram in a PDF file (*see* Fig. 21) and stats in a text file.

```
for (i in c(1:5)) {
        chr = Bdchr$seqnames[i]
        pdf(paste("methstats-",chr,".cpg.pdf", sep=""),
 width=8, height=8, pointsize=12)
        stats = methstats(Bdistachyon.cpg, chrom=chr,
 mcClass='mCG', Nproc=8)
        dev.off()
        sink(paste("methstats-",chr,".cpg.txt", sep=""))
        stats
        sink()
}
```

Below shows the content of methstats-Bd1.cpg.txt that contains the basic stats of CpG methylation profile of chromosome Bd1:

```
$descriptive_stats
     leaf1            leaf2            leaf3            spike1
 Min.    :0.0000    Min.    :0.0000    Min.    :0.0000    Min.
:0.0000
  1st Qu.:0.9000     1st Qu.:0.8890     1st Qu.:0.9000     1st
Qu.:0.8930
 Median :0.9560    Median :0.9520    Median :0.9550    Median
:0.9510
 Mean    :0.8728    Mean    :0.8495    Mean    :0.8693    Mean
:0.8742
  3rd Qu.:1.0000     3rd Qu.:1.0000     3rd Qu.:1.0000     3rd
Qu.:1.0000
 Max.    :1.0000    Max.    :1.0000    Max.    :1.0000    Max.
:1.0000
     spike2           spike3
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.9000   1st Qu.:0.8890
 Median :0.9550   Median :0.9520
 Mean   :0.8747   Mean   :0.8567
 3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000

$correlation_mat
         leaf1     leaf2     leaf3    spike1    spike2    spike3
leaf1   1.0000000 0.6717361 0.6790505 0.6628270 0.6669396
0.6554795
leaf2   0.6717361 1.0000000 0.6756861 0.6399643 0.6527118
0.6718338
leaf3   0.6790505 0.6756861 1.0000000 0.6514059 0.6634676
0.6588023
spike1 0.6628270 0.6399643 0.6514059 1.0000000 0.6636837
0.6463716
spike2 0.6669396 0.6527118 0.6634676 0.6636837 1.0000000
0.6571999
spike3 0.6554795 0.6718338 0.6588023 0.6463716 0.6571999
1.0000000
```

3.8.7 Calculate Differential DNA Methylation

The findDMR function can be used to identify differentially methylated regions (DMRs) of a given BSdataSet object. The Wilcoxon signed rank test is carried out to compare the mC methylation levels between two paired samples, whereas the Kruskal-Wallis test is used when comparing within a group of N samples. The MCClass parameter defines the cytosine context (mCG, mCHG or mCHH) to be considered. The dmrSize parameter defines the number of consecutive mC to be simultaneously considered. The dmrBp parameter defines the maximum number of base pairs containing the dmrSize mC. The ROI parameter can be NULL or given a GRanges object consisting of genomic regions of interest

to confine the search to the specified locations. Run the following command to perform DMR identification.

```
DMR.cpg <- findDMR(object=Bdistachyon.cpg, MCClass='mCG',
dmrSize=10, dmrBp=1000, ROI=NULL)
```

The findDMR function returns a GRanges object. The MethDiff_Perc column stores the percentage methylation difference between the mean methylation of the two sample groups (i.e. leaf and spike) and log2Enrichment column stores the log2 of mean methylation of spike over leaf samples.

```
> head(DMR.cpg,3)
GRanges object with 3 ranges and 3 metadata columns:
      seqnames         ranges strand |    pValue MethDiff_Perc
log2Enrichment
         <Rle>      <IRanges>  <Rle> | <numeric>     <numeric>
<numeric>
  [1]      Bd1 [10586, 10677]      * |     0.118         0.183
0.003
  [2]      Bd1 [11134, 11244]      * |     0.917         3.657
0.055
  [3]      Bd1 [18453, 18505]      * |     0.033          1.28
0.019
  ------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

The consolidateDMRs function is used to select hypermethylated or hypomethylated regions based on the *P*-value (pvThr parameter) and/or the absolute methylation difference percentage thresholds (MethDiff_Thr parameter). The correct parameter defines if the *P*-values are adjusted using the Benjamini-Hochberg multiple testing correction method. DMRs lie within a given distance can be joined by specifying the GAP parameter. The function returns a GRanges object with the mean methylation difference recalculated and the P-values combined using the Fisher's method.

```
DMR_Hyper.cpg <- consolidateDMRs(DmrGR=DMR.cpg, type="hyper",
correct=TRUE, pvThr=0.05, GAP=100)
DMR_Hypo.cpg <- consolidateDMRs(DmrGR=DMR.cpg, type="hypo",
correct=TRUE, pvThr=0.05, GAP=100)
```

Retrieve the number of hypermethylated DMRs with length function:

```
> length(DMR_Hyper.cpg)
[1] 4120
```

```
> head(DMR_Hyper.cpg,3)
GRanges object with 3 ranges and 3 metadata columns:
      seqnames          ranges strand |    pValue MethDiff_Perc
```

```
log2Enrichment
             <Rle>           <IRanges>   <Rle> | <numeric>
<numeric>      <numeric>
  [1]      Bd1 [ 45495,  45578]      * |         0        16.697
0.304
  [2]      Bd1 [139175, 139342]      * |         0        14.467
3.839
  [3]      Bd1 [202519, 203314]      * |         0         17.35
1.878
  ------
   seqinfo: 6 sequences from an unspecified genome; no
seqlengths
```

Retrieve the number of hypomethylated DMRs with length function:

```
> length(DMR_Hypo.cpg)
[1] 3521

> head(DMR_Hypo.cpg,3)
GRanges object with 3 ranges and 3 metadata columns:
     seqnames          ranges strand |    pValue MethDiff_Perc
log2Enrichment
             <Rle>           <IRanges>   <Rle> | <numeric>
<numeric>      <numeric>
  [1]      Bd1 [178034, 178064]      * |         0        -1.823
-0.028
  [2]      Bd1 [306182, 306244]      * |         0        -1.913
-3.258
  [3]      Bd1 [325314, 325435]      * |     0.046        -6.363
-0.191
  ------
   seqinfo: 7 sequences from an unspecified genome; no
seqlengths
```

3.8.8 Plot DNA Methylation Over Genomic Region of Interest

The plotMeth function allows the methylation data of multiple samples to be displayed at a given regions of interest together with the locations of transcripts and additional annotation information. Below I showed two examples whereby the differentially hypermethylated and hypomethylated regions identified in Subheading 3.8.7 are visualized using the plotMeth function.

A differentially hypermethylated region that is more than 1000 bp in length and with a percentage methylation difference greater than 30% in spike tissues is found in chromosome Bd4:

```
> DMR_Hyper.cpg[ranges(DMR_Hyper.cpg)@width > 1000 & element-
Metadata(DMR_Hyper.cpg)$MethDiff_Perc > 30]
GRanges object with 1 range and 3 metadata columns:
     seqnames          ranges strand |    pValue MethDiff_Perc
            <Rle>           <IRanges>   <Rle> | <numeric>
```

```
<numeric>
  [1]     Bd4 [8599804, 8600933]     * |        0       33.595
      log2Enrichment
           <numeric>
  [1]          1.768
```

The region is selected and saved into a Granges object named exHyper.

```
exHyper = DMR_Hyper.cpg[ranges(DMR_Hyper.cpg)@width > 1000 &
elementMetadata(DMR_Hyper.cpg)$MethDiff_Perc > 30]
```

DNA methylation profiling is performed on this genomic region for each of the six sample by calling the profileDNAmetBin function. The genomic region can be divided into several bins (default nbins is 2) and for each bin, the density of cytosines (C/bp), absolute methylation density (mC/bp), and relative methylation density (mC/C) are determined. Executing the following commands to obtain six objects of class GEcollection (collection of genomic regions):

```
# nbins=1 to avoid NA in the calculation
gec.leaf1 <- profileDNAmetBin(GenoRanges=exHyper, Sam-
ple=leaf1.db, mcCLASS='mCG', nbins=1)
gec.leaf2 <- profileDNAmetBin(GenoRanges=exHyper, Sam-
ple=leaf2.db, mcCLASS='mCG', nbins=1)
gec.leaf3 <- profileDNAmetBin(GenoRanges=exHyper, Sam-
ple=leaf3.db, mcCLASS='mCG', nbins=1)
gec.spike1 <- profileDNAmetBin(GenoRanges=exHyper, Sam-
ple=spike1.db, mcCLASS='mCG', nbins=1)
gec.spike2 <- profileDNAmetBin(GenoRanges=exHyper, Sam-
ple=spike2.db, mcCLASS='mCG', nbins=1)
gec.spike3 <- profileDNAmetBin(GenoRanges=exHyper, Sam-
ple=spike3.db, mcCLASS='mCG', nbins=1)
```

Save multiple GEcollection objects into a GElist object:

```
gel <- GElist(gecLeaf1=gec.leaf1, gecLeaf2=gec.leaf2, ge-
cLeaf3=gec.leaf3, gecSpike1=gec.spike1, gecSpike2=gec.
spike2, gecSpike3=gec.spike3)
```

Create a data.frame object containing the "mock" cytoband information of *B. distachyon* Bd21. The data.frame should be substituted with real data for genomes that have known cytoband information from Giemsa staining.

```
bandDF = data.frame(chrom = seqnames(Bdistachyon), chromStart
= 0, chromEnd = seqlengths(Bdistachyon), name = seqnames
(Bdistachyon), gieStain = "gneg")
```

Create variables containing the information required for inputting into the plotMeth function, including the chromosome of the genomic region, the left and right coordinates of the plotted

region, and the maximum absolute methylation level among the six sample.

```
txdb = TxDb.Bdistachyon.JGI.Bd3.1.geneexons
chrom = as.character(seqnames(exHyper)@values)
pos1 = ranges(exHyper)@start - 2000
pos2 = ranges(exHyper)@start+ranges(exHyper)@width + 2000
ylimit = round(max(binmC(gec.leaf1), binmC(gec.leaf2), binmC
(gec.leaf3), binmC(gec.spike1), binmC(gec.spike2), binmC(gec.
spike3), na.rm = TRUE),2)
```

The plotMeth function is called and the figure is saved into a PNG file (*see* Fig. 22).

```
png("plotMeth.hyper.cpg.png", width=2000, height=3000,
res=300)
plotMeth(gel, colors=c(rep("red",3),rep("blue",3)),
datatype=c(rep("mC",6)), yLim=c(rep(ylimit,6)), brmeth=list
(leaf1=leaf1.db, leaf2=leaf2.db, leaf3=leaf3.db, spi-
ke1=spike1.db, spike2=spike2.db, spike3=spike3.db), mcCon-
text="CG", transcriptDB=txdb, chr=chrom, start=pos1,
end=pos2, org=Bdistachyon, ucsc=FALSE, bands=bandDF, annoda-
ta=GRangesList("HYPER"= exHyper))
dev.off()
```

In the second example, a differentially hypomethylated region that is more than 1000 bp in length and with percentage methylation difference greater than 30% in leaf samples is found in chromosome Bd3:

```
> DMR_Hypo.cpg[ranges(DMR_Hypo.cpg)@width > 1000 & elementMe-
tadata(DMR_Hypo.cpg)$MethDiff_Perc < -30]
GRanges object with 1 range and 3 metadata columns:
     seqnames           ranges strand |    pValue MethDiff_Perc
            <Rle>              <IRanges>   <Rle> | <numeric>
<numeric>
  [1]      Bd3 [5658620, 5659769]      * |     0.015        -31.4
     log2Enrichment
          <numeric>
  [1]         -0.756
```

Execute the following commands to plot this hypomethylated region (*see* Fig. 23).

```
exHypo = DMR_Hypo.cpg[ranges(DMR_Hypo.cpg)@width > 1000 &
elementMetadata(DMR_Hypo.cpg)$MethDiff_Perc < -30]

# nbins=1 to avoid NA in the calculation
gec.leaf1 <- profileDNAmetBin(GenoRanges=exHypo, Sam-
ple=leaf1.db, mcCLASS='mCG', nbins=1)
gec.leaf2 <- profileDNAmetBin(GenoRanges=exHypo, Sam-
ple=leaf2.db, mcCLASS='mCG', nbins=1)
gec.leaf3 <- profileDNAmetBin(GenoRanges=exHypo, Sam-
```

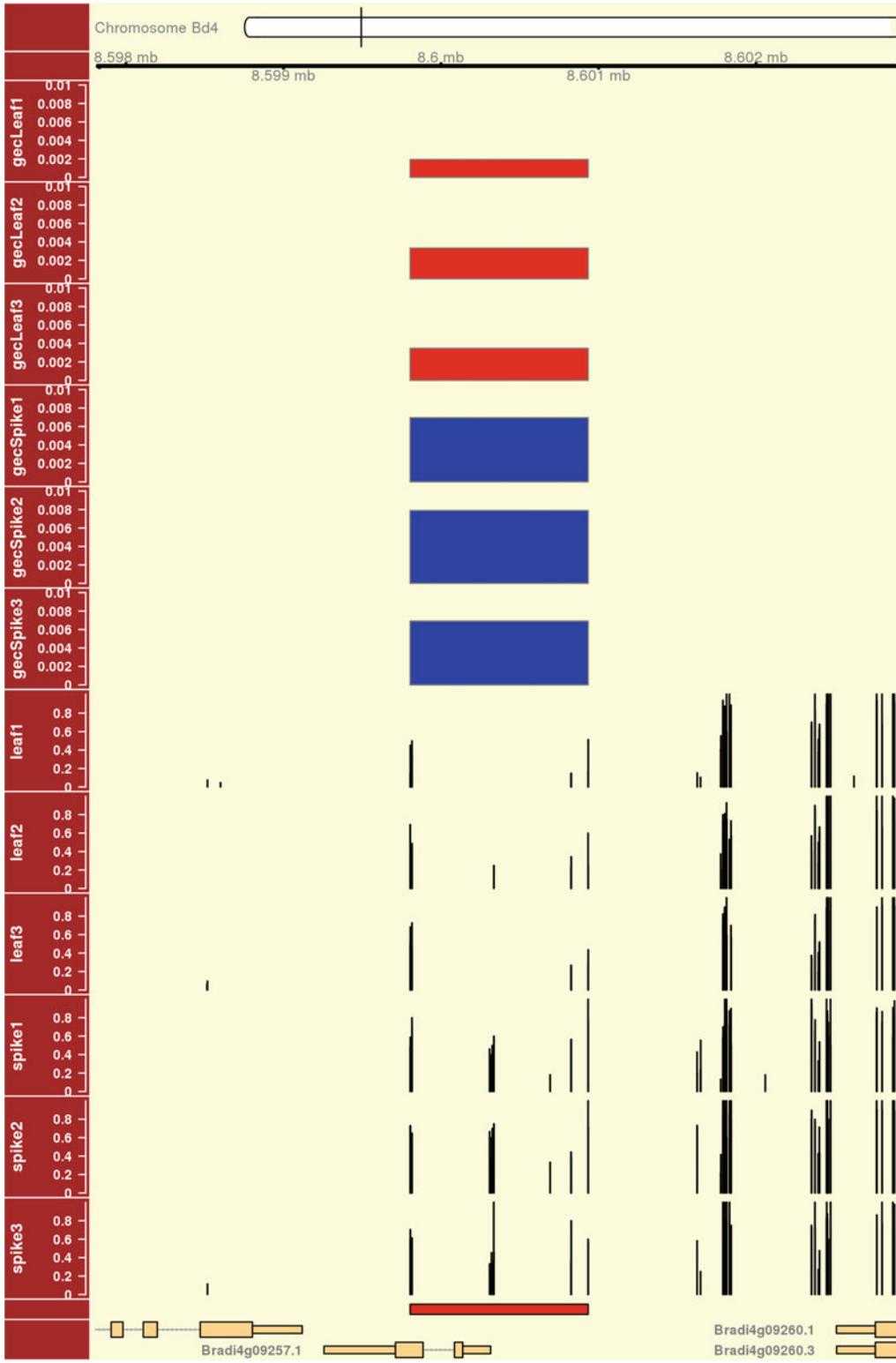

Fig. 22 Genomic view of Bd4:8597804-8602934 that shows a differentially hypermethylated region in spike samples. The hypermethylated region overlaps the promoter and gene body of Bradi4g09257.1

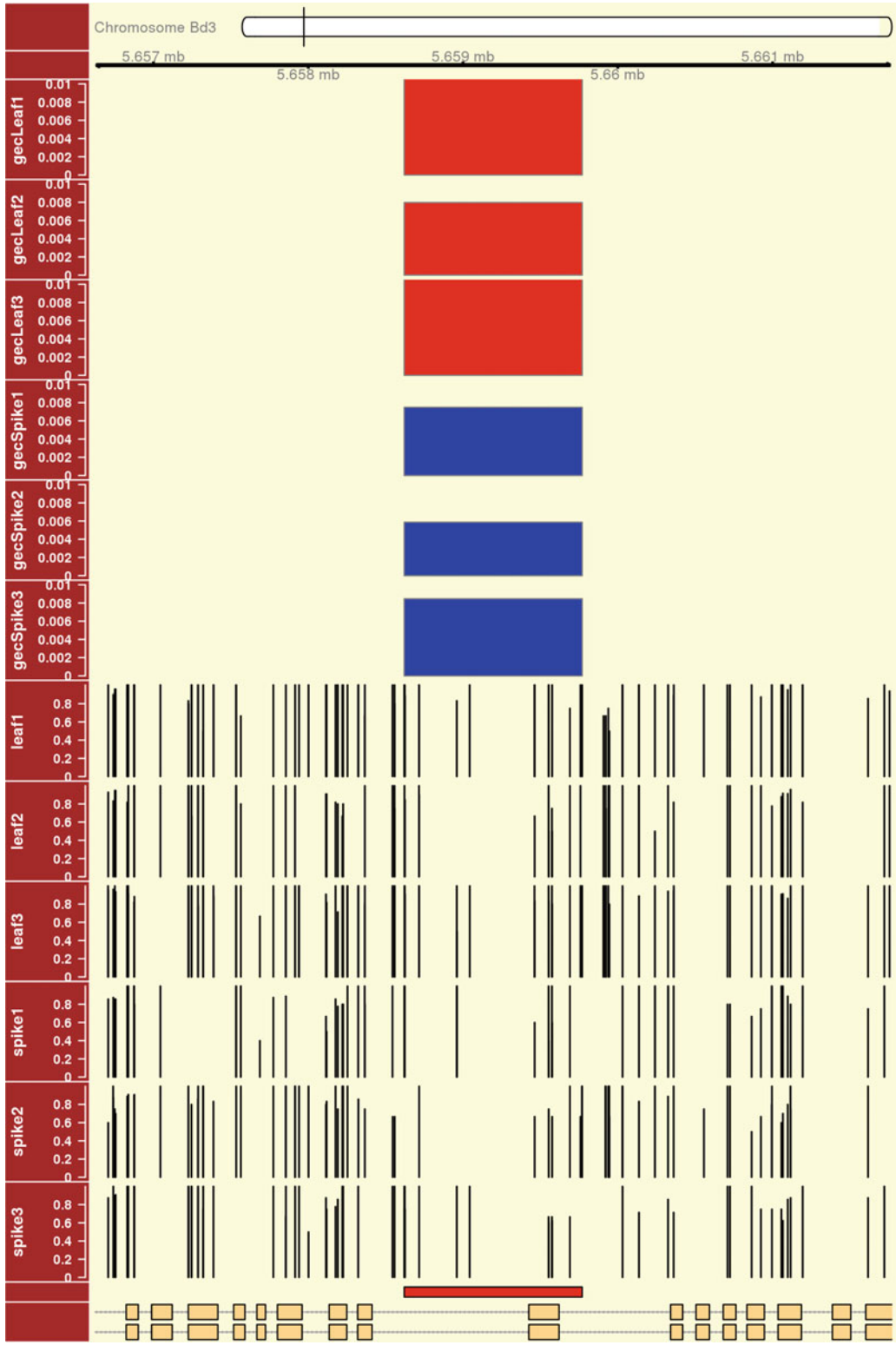

Fig. 23 Genomic view of Bd3:5656620-5661770 that shows a differentially hypomethylated region in spike samples. The hypomethylated region overlaps the part of the gene body of Bradi3g07490.1 and Bradi3g07490.3

```
ple=leaf3.db, mcCLASS='mCG', nbins=1)
gec.spike1 <- profileDNAmetBin(GenoRanges=exHypo, Sam-
ple=spike1.db, mcCLASS='mCG', nbins=1)
gec.spike2 <- profileDNAmetBin(GenoRanges=exHypo, Sam-
ple=spike2.db, mcCLASS='mCG', nbins=1)
gec.spike3 <- profileDNAmetBin(GenoRanges=exHypo, Sam-
ple=spike3.db, mcCLASS='mCG', nbins=1)

gel <- GElist(gecLeaf1=gec.leaf1, gecLeaf2=gec.leaf2, ge-
cLeaf3=gec.leaf3, gecSpike1=gec.spike1, gecSpike2=gec.
spike2, gecSpike3=gec.spike3)

chrom = as.character(seqnames(exHypo)@values)
pos1 = ranges(exHypo)@start - 2000
pos2 = ranges(exHypo)@start+ranges(exHypo)@width + 2000
ylimit = round(max(binmC(gec.leaf1), binmC(gec.leaf2), binmC
(gec.leaf3), binmC(gec.spike1), binmC(gec.spike2), binmC(gec.
spike3), na.rm = TRUE),2)

png("plotMeth.hypo.cpg.png", width=2000, height=3000,
res=300)
plotMeth(gel, colors=c(rep("red",3),rep("blue",3)),
datatype=c(rep("mC",6)), yLim=c(rep(ylimit,6)), brmeth=list
(leaf1=leaf1.db, leaf2=leaf2.db, leaf3=leaf3.db, spi-
ke1=spike1.db, spike2=spike2.db, spike3=spike3.db), mcCon-
text="CG", transcriptDB=txdb, chr=chrom, start=pos1,
end=pos2, org=Bdistachyon, ucsc=FALSE, bands=bandDF, annoda-
ta=GRangesList("HYPO"= exHypo))
dev.off()
```

References

1. University of California D (2012) Brachypodium distachyon (stiff brome) BS-seq. http://www.ncbi.nlm.nih.gov/bioproject/PRJNA182267. Accessed 12 Feb 2016
2. Song Q, Decato B, Hong EE et al (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One 8(12):e81148
3. DOE-JGI (2016) Phytozome v11: Brachypodium distachyon (purple false brome). http://phytozome.jgi.doe.gov/. Accessed 12 Feb 2016
4. Song Q, Decato B, Kessler M et al. (2015) MethPipe manual. https://github.com/smithlabcode/methpipe/tree/master/docs. Accessed 12 Feb 2016
5. Akalin A, Kormaksson M, Li S et al (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13(10):R87
6. Gu Z (2016) EnrichedHeatmap: making enriched heatmaps. R package version 1.1.5. https://github.com/jokergoo/EnrichedHeatmap. Accessed 9 Apr 2016
7. Kishore K, de Pretis S, Lister R et al (2015) methylPipe and compEpiTools: a suite of R packages for the integrative analysis of epigenomics data. BMC Bioinformatics 16:313

Chapter 18

Application of Tissue Culture and Transformation Techniques in Model Species *Brachypodium distachyon*

Bahar Sogutmaz Ozdemir and Hikmet Budak

Abstract

Brachypodium distachyon has recently emerged as a model plant species for the grass family (*Poaceae*) that includes major cereal crops and forage grasses. One of the important traits of a model species is its capacity to be transformed and ease of growing both in tissue culture and in greenhouse conditions. Hence, plant transformation technology is crucial for improvements in agricultural studies, both for the study of new genes and in the production of new transgenic plant species. In this chapter, we review an efficient tissue culture and two different transformation systems for *Brachypodium* using most commonly preferred gene transfer techniques in plant species, microprojectile bombardment method (biolistics) and *Agrobacterium*-mediated transformation.

In plant transformation studies, frequently used explant materials are immature embryos due to their higher transformation efficiencies and regeneration capacity. However, mature embryos are available throughout the year in contrast to immature embryos. We explain a tissue culture protocol for *Brachypodium* using mature embryos with the selected inbred lines from our collection. Embryogenic calluses obtained from mature embryos are used to transform *Brachypodium* with both plant transformation techniques that are revised according to previously studied protocols applied in the grasses, such as applying vacuum infiltration, different wounding effects, modification in inoculation and cocultivation steps or optimization of bombardment parameters.

Key words *Brachypodium distachyon*, Microprojectile bombardment (Biolistics), *Agrobacterium*-mediated transformation, Plant tissue culture, Plant transformation, Mature embryo-derived callus culture

1 Introduction

The improvements in the studies of molecular biology and genetics depend highly on the availability of a suitable model organism. *Brachypodium distachyon* (hereafter *Brachypodium*), a member of *Poaceae* family, has recently emerged as a new model plant for the diverse and economically important group of grasses used as food and feed supply, and herbaceous energy crops. While these crops relatively have complex genomes and unfavorable growth conditions, *Brachypodium* has all the characteristics to be a tractable

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4_18,

model organism such as having a compact genome (~272 Mbp, $2n = 10$), short generation time (12 weeks), five pair of chromosomes ($2n = 10$), the ability to self-pollinate and easy growth under simple environmental conditions [1]. Until recently, rice (*Oryza sativa*) and *Arabidopsis thaliana* has served as model species for these temperate grasses. However, *Brachypodium*, with all its biological and genetic characteristics and phylogenetic position, being closely related to grass family has become a more suitable model plant than the former ones by overcoming all their limitations [2]. Besides all, the annotated genome sequence of *Brachypodium*, the diploid line Bd21, was published [3] and afterwards it became a more powerful tool for agricultural studies.

Increasing the nutritional value and quality of the food or the crop yield are crucial improvements to be assessed in agricultural studies. Therefore, better understanding of agronomic traits is necessary. With the use of extending availability of genome sequences and gene transfer technology, crop improvement studies have accelerated seriously. Plant genomics studies together with transgenic technology offers a wide range of advances in functional gene analysis. So, both for the study of new genes and in the production of new transgenic plant species, plant transformation technology is crucial for improvements in agricultural studies. Establishment of an efficient plant regeneration system is prerequisite for the success of plant transformation applications.

The efficiency of transformation using either *Agrobacterium*-mediated transformation or the microprojectile bombardment method is mainly affected by the genotype, explant type and age, culture conditions, choice of promoter type, and the type of selectable markers used. So the tissue culture conditions before and after transformation and optimization of the parameters of the transformation system are crucial for the success of transformation. Embryogenic callus, due to its higher regeneration capacity, has a great influence on the transformation efficiency. Potential production of embryogenic callus differs according to the species, cultivars of that species and the source of explant and media composition used for callus initiation. For *Brachypodium* species, both mature [4, 5] and immature embryos [1, 4–9] were used to obtain embryogenic callus formation according to the conditions mainly described in the work of Bablak et al. [4]. Tetraploid accessions produced callus at higher percentages compared to diploid ones. Meanwhile, immature embryos were regenerated more effectively with respect to mature ones in overall Scheme [5]. Though immature embryos have good regeneration capacities, mature embryos are available explant sources throughout the year that it speeds up the procedures in genetic transformation studies.

Brachypodium genetic transformation was firstly achieved with the study of Draper et al. [1]. Calluses were formed using immature embryos of different sizes with two different media composition.

Hygromycin-resistant plants from ABR100 accession were developed by microprojectile bombardment. Christiansen et al. [6] selected two diploid (BDR001 and BDR018) and two tetraploid (BDR017 and BDR030) accessions and utilized immature embryos to induce embryogenic callus formation. These tissues were bombarded and higher transformation efficiencies were achieved from tetraploid ones. In 1-year time, they had tested T_0 and T_1 generations and produced T_2 seeds due to short life cycle of *Brachypodium* species, implying the importance of *Brachypodium* as a model species.

For different *Brachypodium* genotypes, several *Agrobacterium*-mediated transformation protocols were developed. Using the *Agrobacterium tumefaciens* strain AGL1 with the vector pDM805 and immature embryos of BDR018 accession, a transformation protocol was optimized to produce BASTA-resistant transgenic plants. It was proposed that T-DNA tagging of diploid *Brachypodium* was possible since the transformation frequency in the study was high enough to achieve in large quantities of transformed lines [7]. In another study, inbred lines of *Brachypodium distachyon* from 27 accessions were developed by applying single seed descent for three or more generations and five of them were found to be diploid that presented different morphological characteristics such as in vernalization requirement [5]. Bd21 was found to have the fastest generation time when compared to others and became a model inbred line. The embryogenic callus formation varied greatly among the genotypes and also the regeneration percentages were in varying rate, being affected both by the genotype and the explant type used for callus induction. The 10 of the 19 lines were transformed *via Agrobacterium*-mediated transformation method using both mature and immature embryos. Super-virulent strain AGL1 was used with three DNA constructs conferring hygromycin resistance. Bd17-2, hexaploid accession, had the highest average transformation efficiency. *Agrobacterium*-mediated transformation of the inbred line Bd21-3 was performed [8] using AGL1 with seven DNA constructs. Immature embryos were transformed with higher transformation efficiencies ranging from 10 to 41% with respect to Bd21 line. The inbred line Bd21 was transformed using *Agrobacterium*-mediated gene transfer system and immature embryos as the source of compact embryogenic callus formation by applying hygromycin selection for T-DNA insertional mutagenesis [9]. The transformation efficiency was reported to be 17% in this study. GFP expression was observed in both T_0 and T_1 progenies. T-DNA insertions and flanking sequence tags (FSTs) produced in this study well define the arguments about the use of model species and transgenic technology together with the functional genomics studies. After this study, they published a protocol for transformation of Bd21 line with *Agrobacterium* method [10]. This standard community line of *Brachypodium* was also

transformed with the *Agrobacterium* AGL1 strain using two binary vectors and immature embryos as explant source. It was shown that high selection pressure was applied using pCAUGH for transformation to save time and labor [11].

Three hydrolase genes were transiently expressed in *Brachypodium* Bd21 line using *Agrobacterium*-mediated transformation of mature embryos [12] and this could serve for further studies in plant cell wall modification. Artificial microRNAs designed for silencing the two candidate genes encoding for CAD and COMT enzymes that may cause altered lignin composition and improved digestibility [13]. Embryogenic calluses of *Brachypodium* Bd21-3 line were transformed by *Agrobacterium* method according to the protocol of Vogel and Hill [8]. These recent transformation studies show the power of *Brachypodium* to serve as a basis for research in biofuel production.

In this protocol, we explain a tissue culture protocol for *Brachypodium* using mature embryos with the selected inbred lines exhibiting different ploidy levels from our collection sampled from diverse geographic regions of Turkey. Embryogenic calluses obtained from mature embryos with a simple and efficient tissue culture protocol were used to transform *Brachypodium* with both biolistics and *Agrobacterium*-mediated transformation that were revised and modified according to previously studied protocols applied in the grasses, such as applying vacuum infiltration, different wounding effects, modification in inoculation and cocultivation steps or optimization of bombardment parameters.

2 Materials

2.1 Plant and DNA Material

1. Plant Material: A total of 146 inbred lines were created from 1101 *Brachypodium* individuals representing diverse geographic regions of Turkey. From this collection, three different *Brachypodium distachyon* genotypes (BdTR4, BdTR6, and BdTR13) showing two different ploidy levels (diploid and tetraploid) were used as plant material for explant source for tissue culture, as well as plant transformation studies. These genotypes were selected according to their morphological traits and growing behavior recorded [14].
2. Plasmid for Biolistic Transformation: The plasmid pCAMBIA1301 (CAMBIA, Canberra, Australia) was used. It carried the genes, β-glucuronidase uidA (*GUS*) gene as reporter and the hygromycin phosphotransferase (*hpt*) for plant selection, which were both driven by the Cauliflower mosaic Virus "35S" (CaMV35S) promoter.
3. Bacterial Strains and Plasmids for *Agrobacterium*-mediated Transformation: Three different *Agrobacterium tumefaciens*

strains AGL1, EHA105 and LBA4404 were used. These strains were chosen according to their transformation efficiency and frequency of usage in monocot systems [15, 16]. AGL1 strain carried the plasmids pAL154 and pAL156. Plasmid pAL156, helper plasmid, contained the *GUS* gene as the reporter and the *bar* gene conferring resistance to glufosinate ammonium based herbicides like PPT (phosphinothricin) for plant selection. Both genes were under the control of *ubi* (ubiquitin) promoter. Kanamycin and carbenicillin resistance genes were integrated for bacterial selection. EHA 105 and LBA4404 strains carried the plasmid pGUSINT that contained the *GUS* gene under the control of CaMV35S promoter. For bacterial and plant selection, they had the *nptII* (neomycin phosphotransferase) gene for kanamycin resistance.

2.2 Culture Media

All media should be autoclaved at 121 °C for 20 min and stored at 4 °C until used.

1. Callus Induction Medium (CIM): 4.43 g/l MS [17] basal salt medium including vitamins supplemented with 30 g/l sucrose and specified concentrations of plant growth regulator (2,4-D; 2,4-Dichlorophenoxyacetic acid) as listed below and solidified with 8 g/l plant agar, pH 6.0 (*see* **Note 1**).
   Concentrations of plant growth regulators supplemented for each media:

   CIM1: 1 mg/l 2,4-D (BdTR13).

   CIM2: 3 mg/l 2,4-D (BdTR6).

   CIM3: 5 mg/l 2,4-D (BdTR4).
2. Regeneration Medium (RM): 4.43 g/l MS basal salt medium including vitamins supplemented with 30 g/l sucrose and solidified with 8 g/l plant agar, pH 6.0.
3. LB medium: 10 g/l Bacto tryptone, 5 g/l yeast extract, and 10 g/l NaCl, pH 7.0.
4. LB agar (LA) medium: Prepare LB medium and add 15 g/l Bacto agar before autoclaving.
5. MGL medium: 5 g/l mannitol, 2.5 g/l yeast extract, 1 g/l glutamic acid, 0.1 g/l $MgSO_4.7H_2O$, 0.25 g/l KH_2PO_4, 0.25 g/l NaCl, and 5 g/l Bacto tryptone, pH 7.0 and supplemented with 1 μg/l filter-sterilized biotin after autoclaving.
6. YEB medium: 13.5 g/l nutrient broth, 1 g/l yeast extract, 5 g/l sucrose, and 2 mM $MgSO_4{\cdot}7H_2O$, pH 7.2.
7. YEB agar medium: Prepare YEB medium and add 15 g/l Bacto agar before autoclaving.
8. YEB-MES medium: YEB medium containing at a final concentration of 10 mM MES [2-(N-morpholino) ethanesulfonic

acid], pH 5.6. MES might be used filter-sterilized and added to the medium after autoclaving.

9. MMA medium: 4.3 g/l MS basal salt medium, 20 g/l sucrose, and 10 mM MES, pH 5.6.
10. MMD medium: 4.4 g/l MS basal salt medium, 10 mM MES, 30 g/l sucrose, 1–5 mg/l 2,4-D (depending on optimized 2,4-D concentrations used in tissue culture media for each genotype), 8 g/l plant agar, pH 5.6 and supplemented with filter-sterile 100 mg/l ascorbic acid and 200 μM acetosyringone after autoclaving.

2.3 Buffers and Stock Solutions

1. 2,4-Dichlorophenoxyacetic acid (2,4-D) (5 mg/ml stock solution): Dissolve in 200-proof ethanol and bring to volume with double distilled water. Store as aliquots at –20 °C.
2. GUS staining (X-gluc) solution: 1 ml (0.1 M) EDTA, 1 ml (5 mM) potassium ferricyanide, 1 ml (5 mM) potassium ferrocyanide, 5.0 ml sodium phosphate buffer (%61 (0.1 M) Na_2HPO_4 and %39 (0.1 M) NaH_2PO_4, pH 7.0), 100 μl (0.1%) Triton X-100, and 0.3 mg/ml X-gluc (dissolved in DMF until transparent) and with sterile distilled water add up the volume to 10 ml.
3. GUS fixative solution: 10% formaldehyde, 20% ethanol, and 5% acetic acid.
4. CTAB (hexadecyl-trimethyl-ammonium bromide) Extraction Buffer (2%): Dissolve 2 g CTAB with 10 ml (1 M) Tris–HCl, add 4 ml (0.5 M) EDTA and 28 ml (5 M) NaCl. Bring the volume to 100 ml with sterile double distilled water. Heat the solution in microwave oven to dissolve CTAB if needed. Keep the solution away from light and store at room temperature. The pH of the solutions used in CTAB buffer should be adjusted to pH 8.0 and each should be autoclaved before use.
5. TE (Tris–EDTA) Buffer: Mix 100 μl (1 M) Tris and 20 μl (0.5 M) EDTA and bring the volume to 10 ml with distilled water, pH 8.0. Autoclave and store at room temperature.
6. 1% agarose gel: Weigh 1 g agarose, put in a flask along with 100 ml 1× TBE. Microwave 1–3 min until the agarose is completely dissolved. Cool down agarose solution for 3–5 min and add 3 μl ethidium bromide from stock solution.
7. Ethidium bromide (EtBr) staining solution (5 mg/ml): Put first 20 ml distilled water in a falcon tube, then add 0.1 g EtBr, mix with a wood stick. Use a mask during preparation of the stock solution since it is a mutagen and cancer-suspect agent. Store at room temperature in a dark bottle or cover the tube with aluminum foil.

8. 10× TBE (Tris-Borate-EDTA) Buffer: 108 g/l Tris base, 55 g/l Boric acid, 40 ml/l (0.5 M EDTA, pH 8.0), pH 8.0. Autoclave and store at room temperature.
9. Acetosyringone (200 mM stock solution): Dissolve 390 mg acetosyringone in 10 ml (70%) ethanol. Filter-sterilize and store as aliquots in microcentrifuge tubes at −20 °C. Do not refreeze again and use fresh every time needed.
10. Ascorbic acid (100 mg/ml stock solution): Dissolve 100 mg ascorbic acid in 1 ml sterile double distilled water and store at 4 °C.
11. Antibiotic stock solutions: 25 mg/ml hygromycin, 50 mg/ml kanamycin, 100 mg/ml carbenicillin, 30 mg/ml rifampicin, 250 mg/ml streptomycin, 250 mg/ml cefotaxime. Store at −20 °C (*see* **Note 2**).
12. PPT (DL-phosphinothricin) (5 mg/ml stock solution): Dissolve 0.05 g PPT in 10 ml sterile double distilled water. Store at 4 °C.
13. Depurinating buffer: 250 mM HCl.
14. Denaturation buffer: 1.5 M NaCl and 0.5 M NaOH.
15. Neutralization buffer: 1.5 M NaCl and 0.5 M Tris–HCl, pH 7.5.
16. Transfer buffer: 10× SSC.
17. 10× SSC (sodium chloride–sodium citrate solution): 0.3 M Na_3citrate and 3 M NaCl, completed to 2 l, pH 7.0.
18. Low stringency buffer I: 2× SSC and 1% SDS (Sodium dodecyl sulfate).
19. Low stringency buffer II: 1× SSC and 1% SDS.

2.4 Molecular Biology and Biolistic Particle Delivery System Kits

1. Molecular biology kits: DNeasy Plant Mini Kit (QIAGEN-69106), Genopure Plasmid Midi Kit (ROCHE-3143414001), QIAprep Spin Miniprep Kit (QIAGEN-27106), QIAquick Gel Extraction Kit, (QIAGEN-28704), and DIG High prime DNA labeling and detection kit (ROCHE-11093657910).
2. Biolistic particle delivery system kits for PDS-100/He and Hepta Systems (Bio-Rad):
   1.0 μm and 1.6 μm Gold Microcarriers, Macrocarriers (65-2335), Stopping Screens (165-2336), Macrocarrier Holders (165-2322), 650 psi Rupture Disks (165-2327), 900 psi Rupture Disks (165-2328), and 1100 psi Rupture Disks (165-2329).

2.5 Equipments

1. Biolistic device (PDS-1000/He Particle Delivery System, Bio-Rad, USA).
2. Camera (Connected to a stereomicroscope).

3. Centrifuges (microfuge, microcentrifuge, and ultracentrifuge).
4. Gel electrophoresis system with power supply.
5. Gel documentation.
6. Heating block.
7. Hybridization oven.
8. Illuminator (light source for microscope).
9. Incubator.
10. Laminar flow (or sterile cabinet).
11. Liquid nitrogen tank.
12. Orbital shaker.
13. Plant growth chamber (or plant growth room).
14. Shaking incubator.
15. Spectrophotometer.
16. Stereomicroscope.
17. Thermocycler.
18. Tissue lyser (or homogenization with mortar and pestle).
19. UV transilluminator.
20. Vacuum pump (Connected to biolistic device).
21. Water bath.

3 Methods

3.1 Growth Conditions of Brachypodium Plants in the Greenhouse

1. Collect the mature *Brachypodium* seeds. Stratify at 4 °C between wetted filter papers in petri plates and keep in the dark for 7–10 days.
2. After cold treatment, put the seeds in petri plates under light at room temperature for 5 days.
3. Transfer the germinated seeds first to peat-soil mixture in the viols and grow them until the emergence of third leaf, and then transfer the plantlets into plastic pots. Grow the plants under a controlled environment (16/8 h light/dark photoperiod at 22/25 °C, relative humidity 60–70%, and a photosynthetic photon flux of 320 μmol/m^2/s at canopy height provided by fluorescent lamps) in the greenhouse.
4. Treat the soil with 200 mg/kg N ($Ca(NO_3)_2$), 100 mg/kg P (KH_2PO_4), 20 mg/kg S (K_2SO_4), 5 mg/kg Fe (Fe-EDTA) and 2.5 mg/kg Zn ($ZnSO_4$) for every 20 days to supply basal fertilization.
5. Cover the seed heads during anthesis to prevent cross-pollination in case it persists (Fig. 1).

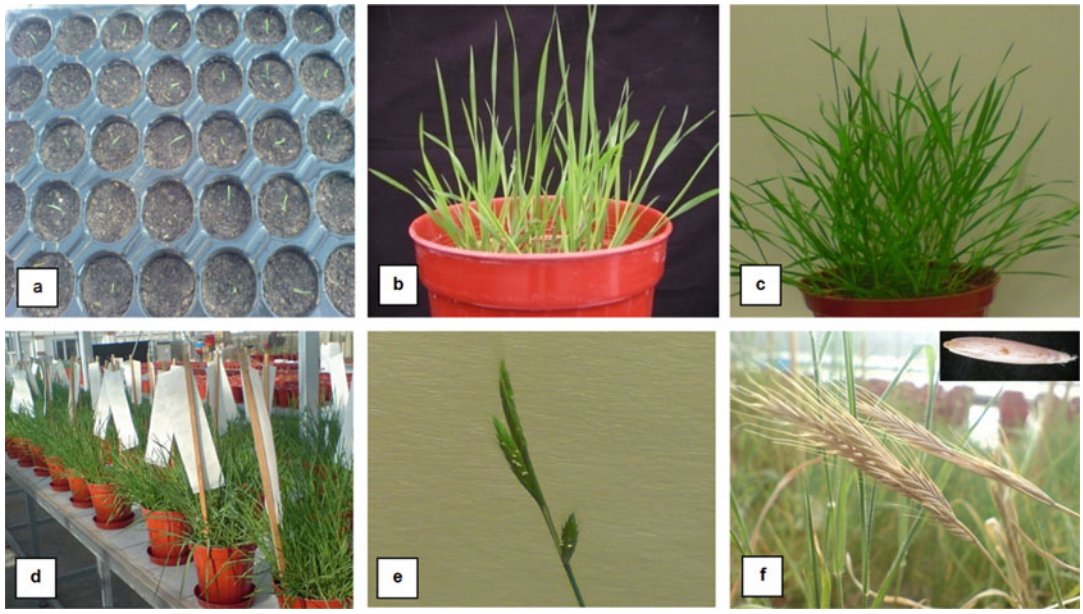

Fig. 1 *Brachypodium distachyon* grown in the greenhouse (**a–f**). Different developmental stages of growth

3.2 Tissue Culture Conditions of Brachypodium

Carry out all the procedures in the laminar flow at room temperature unless otherwise stated. All equipment should be autoclaved before use.

3.2.1 Seed Surface Sterilization

1. Removal the palea and lemma of *Brachypodium* mature seeds.
2. Put the seeds in 70% (v/v) ethanol for 5 min and shake occasionally. Then, wash three times with sterile distilled water.
3. Treat with commercial bleach (53% NaOCl) and 1–2 drops of Tween-20 for 20 min by shaking occasionally. Then, rinse 5 times with sterile distilled water.

3.2.2 Explant Choice, Preparation, Callus Induction, and Regeneration

As the explant source, stem (mesocotyl tissue), root tip, leaf segment, and mature embryos of *Brachypodium* were used for the preliminary optimization of callus initiation (*see* **Note 3**). For callus induction of all explant types; two different carbohydrate sources (maltose or sucrose) at an amount of 30 g/l, different auxin types (2.4-D, IAA, and NAA) at three concentration levels (1, 3, and 5 mg/l), and two levels of cytokinin (BAP; 0.0 and 0.5 mg/l) were applied. Each experiment was set with four replicates. However, efficient embryogenic callus formation was obtained from mature embryos. Our results also showed that each genotype needed different concentrations of auxin hormone (2,4-D) for callus formation and usage of cytokinin caused necrosis problem. Mature seeds of BdTR13, BdTR6, and BdTR4 were induced to make callus formation using different media as CIM1, CIM2, and CIM3, respectively.

1. Excise mature embryos after imbibition of surface sterilized seeds in sterile distilled water for 1–2 h at 33 °C in water bath and put the embryos (20 embryos/petri plate) onto specified callus induction media (CIM1, CIM2, or CIM3).
2. Seeds can also be directly placed onto the callus induction media by skipping the embryo excision step. When the callus is grown sufficient enough to handle, remove the callus tissue from the seed and transfer to fresh callus induction media (*see* **Note 4**).
3. Keep the cultures in dark at 25 °C ± 1 for callus initiation and subculture to fresh media every 2–3 weeks. Select the embryogenic calluses formed and break into pieces before transferring to new media (*see* **Note 5**).
4. After 2–3 months of callus initiation, put the calluses onto regeneration media (RM) in sterile magenta boxes and incubate under a 16 h light–8 h dark photoperiod at 25 °C ± 1 for plant regeneration (Fig. 2). Subculture to fresh media every 2–3 weeks. Both shoot and root development can be established using only RM media.

3.3 Genetic transformation of Brachypodium with Biolistic® PDS-1000/He Device

3.3.1 Preparation of Plant Materials for Bombardment Process

1. Prior to bombardment, arrange 6-week-old embryogenic calluses in a circle of 2.5 cm-diameter at the center of the petri plates containing fresh CIM and incubate for 1 day until the bombardment process starts (Fig. 3).
2. Use common bean (*Phaseolus vulgaris* L.) cotyledons as control explants for the system check.
3. For surface sterilization, put the bean seeds in 70% (v/v) ethanol for 3 min, wash two times with sterile distilled water, treat with commercial bleach (53% NaOCl) for 15 min and rinse again for three times with sterile distilled water.
4. Place the seeds in regeneration medium in magenta boxes under light at 25 °C ± 1.
5. Excise the cotyledons after 10 days and place them abaxial side up within a 2.5 cm circle in the center of the petri plates.

3.3.2 Chemically Competent Cell Preparation

DH5α strain of *E. coli* was used for transformation.

1. Plate the cells on LB agar (LA) and grow overnight at 37 °C by placing the petri plates upside down in the incubator.
2. Take a single colony, inoculate in 5 ml LB and grow overnight in a shaking incubator with 250 rpm at 37 °C.
3. Inoculate 1 ml of this culture into 50 ml LB and grow until OD_{600} reached 0.3–0.4 under same conditions in a shaking incubator.

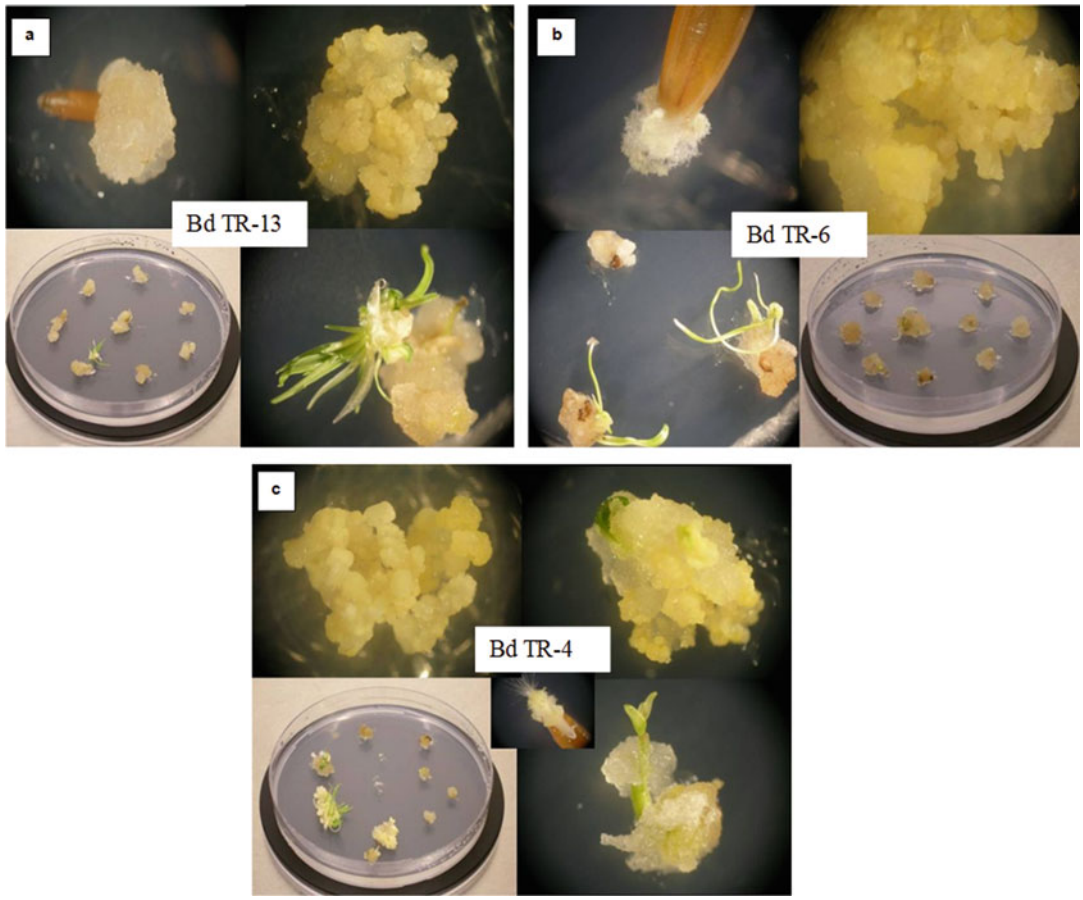

Fig. 2 Illustration of callus formation and regeneration from mature embryos of selected genotypes (**a**) BdTR13; (**b**) BdTR6; (**c**) BdTR4 under optimized conditions. Embryogenic callus formation could be distinguished with its yellowish color and nontranslucent form

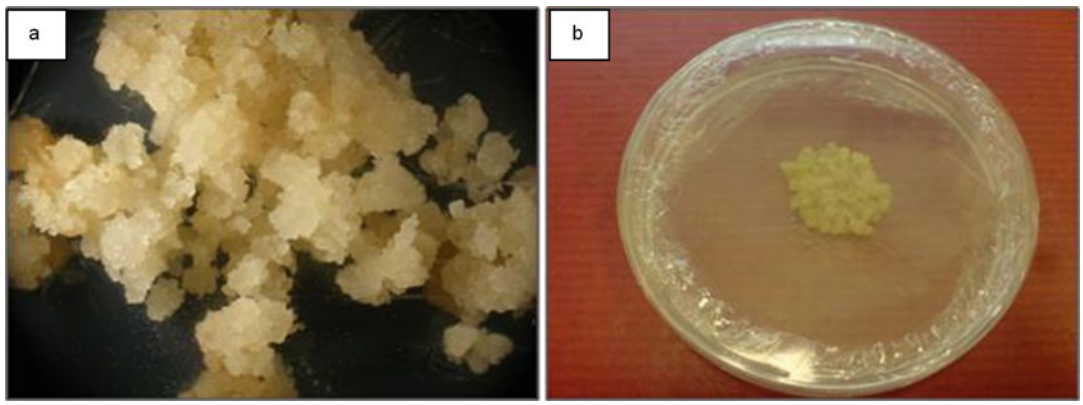

Fig. 3 (**a**) Callus formation from mature embryos of *Brachypodium distachyon*, and (**b**) arrangement of these calluses in the middle of petri plates prior to bombardment

4. Centrifuge the culture at 4100 × *g* for 15 min, resuspend in 25 ml ice-cold 0.1 M $CaCl_2$ and keep on ice for 15 min.
5. Centrifuge the suspension again at 4100 × *g* for 15 min, resuspend in 3.3 ml (1/15 vol) 0.1 M $CaCl_2$—15% glycerol and keep on ice overnight.
6. Freeze the suspension culture as aliquots in liquid nitrogen and store at −80 °C.

3.3.3 Plasmid Transformation and Isolation

1. Mix briefly 1 μl of vector and 200 μl of DH5α competent cells (thawed on ice) in a microcentrifuge tube.
2. Keep the mixture on ice for 20–30 min, then place at 42 °C for 90 s, and lastly keep on ice for 1–2 min.
3. Add 800 μl LB medium onto the transformed cells and keep the tube at 37 °C for 45 min.
4. Spread 50 μl of the culture over LA plates supplemented with 50 mg/l kanamycin antibiotic and keep the plates at 37 °C overnight in an incubator.
5. Inoculate single colonies in 5 ml LB media with 50 mg/l kanamycin antibiotic and again grow the liquid culture overnight in a shaking incubator at 37 °C with 250 rpm.
6. Mix 800 μl of the culture with 200 μl sterilized 87% glycerol.
7. Freeze the prepared stock cultures as aliquots in liquid nitrogen and store at −80 °C until use.
8. Isolate the plasmid from overnight grown culture using a proper plasmid isolation kit by carrying out the procedures according to manufacturer's instructions.
9. Determine the concentration using a spectrophotometer.
10. Confirm the availability of isolated plasmids by enzyme digestion. (In this study, 1 μg pCAMBIA1301 plasmid is digested with Xho I enzyme at 37 °C for 2–3 h).
11. Check the restricted fragments using an agarose gel electrophoresis on a 1% agarose gel in 1× TBE buffer.

3.3.4 Bombardment of Explants with DNA-Coated Gold Particles

Two different sizes of gold particles (1.0 and 1.6 μm in average diameter) were used as microprojectiles (microcarriers). Calluses and cotyledons were bombarded under a partial vacuum of 27″ Hg pressure with three different bombardment pressures (650, 900 and 1100 psi rupture disks) and two sample plate distances (6 and 9 cm) (*see* **Note 6**).

1. Prepare 50 μl gold suspension for every 4–5 shots.
2. Weigh required amount of gold particles on an analytical balance.
3. Add 1 ml of 100% ethanol for every 60 mg of gold particle weighed.

4. Vortex for 1–2 min, centrifuge at 9200 × *g* for 1 min at 4 °C and remove the supernatant. Repeat these steps three times to fully clean the gold particles.
5. Replace the ethanol with 1 ml of sterile distilled water and rinse two times by vortexing for 1–2 min and centrifuge at 9200 × *g* for 1 min at 4 °C.
6. Resuspend the pellet in 1 ml of 50% sterile glycerol.
7. For every 50 µl gold suspension, add 6 µl DNA (1 µg/1 µl), 50 µl of 2.5 M $CaCl_2$, and 20 µl of 0.1 M spermidine in this order with continuous vortexing (*see* **Note 7**).
8. Centrifuge the suspension at 9200 × *g* for 10 s at 4 °C and discard the supernatant.
9. Wash the pellet with 250 µl of 100% ethanol, centrifuge at 9200 × *g* for 10 s at 4 °C and resuspend in 60 µl of 100% ethanol.
10. Keep the suspension at 4 °C until bombardment.
11. Before each bombardment, clean vacuum chamber and its components (stopping plate, sample plate and sample chamber) with ethanol.
12. Sterilize the stopping screens, and macroprojectiles and their holders with ethanol; and rupture (pressure) disks with isopropanol.
13. Place the macroprojectiles in their holders, load each macroprojectile with 8 µl of suspension and allow drying under vacuum.
14. Carry out all steps of bombardment as described in manufacturer's protocol.

3.3.5 Assay of Transient Gene Expression

Transient expression of the GUS gene was detected by "Histochemical GUS Staining" method [18]. The GUS gene products (β-glucuronidase enzyme) produced blue dyes (blue spots) at the site of enzyme activity when the embryos supplied with the chromogenic substrate "X-gluc" (5,bromo-4,chloro-3,indolyl β-D-glucuronic acid) in the presence of atmospheric oxygen.

Highest transient expression was achieved from tissues bombarded with 1.6 µm sized gold particles at a bombardment condition of 1100 psi-6 cm. However, regenerated tissues showed high percentage of albino shoot formation. Highest regeneration percentage was achieved from the hygromycin resistant plants (BdTR4 and BdTR13) bombarded at 650 psi-9 cm and 1100 psi-6 cm with 1.0 µm sized gold particles.

1. After bombardment, incubate the explants (calluses and cotyledons) for 48 h at 25 °C in dark.
2. Put half the number of calluses and all cotyledons bombarded into proper tubes.

3. Wash the explants with sterile distilled water to remove the agar particles from the explant surface.
4. Apply the GUS assay by adding X-gluc solution to the tubes until it covers the surface of the explants. Then, incubate at 37 °C in darkness for 24–48 h.
5. Keep explants in absolute ethanol or GUS fixative solution after incubation.
6. Record the GUS expression by counting the blue spots (putative transformed cells or cell aggregates) under microscope.
7. At the same time, culture the rest of the calluses that are not subjected to assay in new callus induction medium (CIM) and keep at dark for further 4–5 day. Then, transfer them to new media for selection.

3.3.6 Selection and Regeneration of the Putative Transformants

1. For the selection of the transformed cells, transfer the calluses to CIM supplemented with 15 mg/l hygromycin antibiotic and keep them in dark at 25 °C ± 1 for 2 weeks.
2. Transfer the calluses that show resistance to the hygromycin-B antibiotic to RM supplemented with 30 mg/l hygromycin in magenta boxes. Apply antibiotic selection for further 1.5–2 months (*see* **Note 8**).
3. Finalize the selection, transfer the resistant plantlets to magenta boxes containing RM and continue to grow under previously specified conditions. Subculture to fresh media every 2–3 weeks.
4. Regenerate and grow the plantlets in growth room (or chamber) under controlled environment (16/8-h (light/dark) photoperiod at 25 °C ± 1 with 60–70% relative humidity).

Acclimatization:

1. Take the regenerated plants out of the media after shoots and roots are fully developed.
2. Rinse the roots under tap water and transfer to peat–soil (1:1) mixture in pots.
3. Continue to keep them in growth room under specified environmental conditions by exposing high humidity for 10 days with gradual decrease from 90% to 70% (*see* **Note 9**).

3.3.7 Molecular Analysis of the Transient Transgene Integration

1. Isolate genomic DNA from fresh leaves of putative transgenic plants by a commercial plant DNA isolation kit and apply the manufacturer's protocol. (You may also use the CTAB protocol explained in Subheading 3.3.8 rather than using a genomic DNA isolation kit.)

2. Determine the DNA concentration and purity with a spectrophotometer to check DNA "purity". Expected value (ratio) for absorbance at 280/260 nm is ~1.8. Preferably, check whether the isolated DNA is intact by running on 1% agarose gel electrophoresis in 1× TBE buffer stained with EtBr.
3. Carry out PCR amplification to detect ~195 bp region of the CaMV35S promoter by using the primers CaMV35SF (5′ GCTCCTACAAATGCCATCA3′), CaMV35SR (5′'GATAGTG GGATTGTGCGTCA3′) under the following conditions: initial denaturation at 94 °C for 5 min; 35 cycles at 94 °C for 1 min, 55 °C for 45 s and 72 °C for 1 min; and final extension at 72 °C for 10 min. PCR reaction in a final volume of 25 μl: 100 ng genomic DNA (or 50 ng plasmid DNA for the positive control), 0.15 mM dNTPs, 2.0 mM $MgCl_2$, 1× PCR Buffer, 0.4 μM of each primer, and 0.025 units/μl of Taq DNA polymerase as the final concentrations. Include always two negative controls (PCR without any DNA; PCR with wild-type DNA) besides a positive control (Fig. 4).
4. Resolve PCR products by electrophoresis and visualize on 1% agarose gel in 1× TBE buffer stained with EtBr.
5. Choose the samples that amplified the desired PCR product and use for southern blotting.

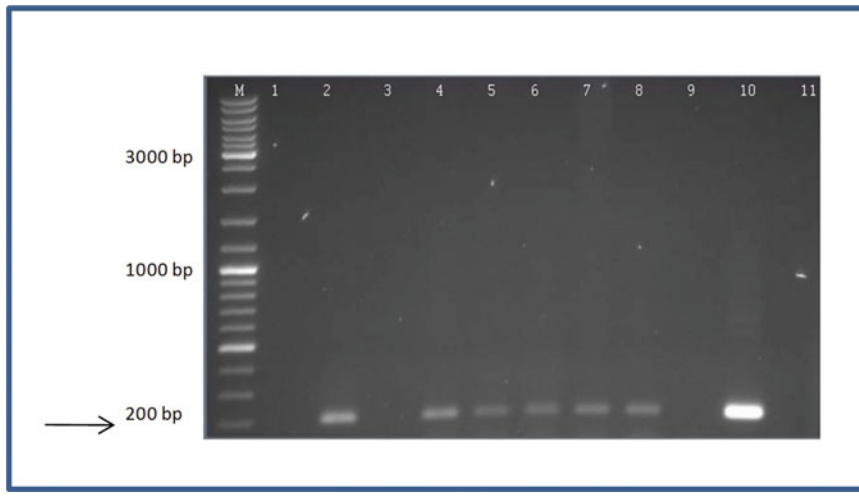

Fig. 4 PCR amplifications of CaMV35S promoter region in five different transformed lines of *Brachypodium distachyon,* visualized with ethidium-bromide stained 1% agarose gel. Arrow indicates the expected PCR product. *Lane M*: molecular marker—Gene Ruler ™ DNA ladder mix; *1*: wild type; *2*: positive control; *3*: negative control; *Lanes 4–8*: hygromycin resistant plants; *9*: negative control; *10*: positive control; *11*: wild type

3.3.8 Southern Blot Analysis for the Confirmation of Stable Transgene Integration

We isolated the genomic DNA of transgenic and nontransgenic plants according to the protocol of Doyle and Doyle [19] with some modifications in order to achieve high concentrations of DNA for southern blot analysis.

1. Homogenize 0.5 g leaf samples with liquid nitrogen until homogenous with sterile mortar and pestle or tissue lyser.
2. Put the powder in 2 ml microcentrifuge tube. Add 500 μl CTAB extraction buffer and add 50 μl β-mercaptoethanol under fume hood. (Store β-ME in dark bottle at 4 °C). Then, add 25 μl (10 mg/ml) add proteinase K.
3. Incubate the tube in water bath at 65 °C for 60 min. Cool the tube briefly on bench for 10 min.
4. Add 500 μl chloroform–isoamyl alcohol (24:1) to tube and mix by vortex.
5. Centrifuge at 18,000 × *g* for 15 min at 4 °C. Collect the supernatant with wide broad tip and transfer into clean tube.
6. Add 500 μl ice-cold isopropanol. Incubate at −20 °C overnight (at least 30 min and up to 3 days).
7. Centrifuge at 18,000 × *g* for 10 min at 4 °C. Pour of the top liquid and wash the pellet twice with 70% cold ethanol.
8. Dry the pellet under laminar flow until it becomes transparent.
9. Dissolve the pellet (DNA) in 50 μl TE (Tris–EDTA) or sterile distilled water.
10. At the end of DNA isolation, treat genomic DNA with 5 μl (10 mg/ml) RNase at 37 °C for 30 min.
11. Precipitate DNA by adding 2× (100%) ethanol and 0.1× (2 M) NaCl onto 1× volume of DNA. Incubate the mixture at −80 °C for 30 min.
12. Centrifuge at 18,000 × *g* for 10 min and wash the pellet with 80% ethanol.
13. Centrifuge again at 18,000 × *g* for 10 min and dry the pellet under laminar flow. Check the translucency of the pellet whether it is fully dried out.
14. Resuspend the pellet (DNA) in TE or sterile distilled water.
15. Determine the DNA concentrations using a spectrophotometer and check by gel electrophoresis as described in Subheading 3.3.7.
16. Digest 10–25 μg of genomic DNA (transgenic and wild type) and 25 pg of plasmid DNA (positive control) overnight with EcoRI restriction enzyme at 37 °C. For 25 μl reaction mixture, put 2.5 μl (10×) EcoRI Buffer), 0.5 μl EcoRI enzyme, and DNA, and add up to with sterile double distilled water.

17. Denature the digested DNA samples at 95 °C for 5 min, leave on ice for 5 min and run on 1% agarose gel containing 0.2 μg final concentration of EtBr in 1× TBE buffer at 25 V for 4 h at 4 °C.
18. After visualization of the digested DNA, rinse the gel with distilled water.
19. Wash the gel with 100 ml depurinating buffer for 10 min and then with denaturation buffer for 25 min.
20. Neutralize the gel finally in neutralization buffer for 30 min. Wash the gel with double distilled water after each buffer treatment.
21. Place the gel in transfer buffer (10× SSC) for 20 min at room temperature.
22. Transfer DNA by downward capillary blotting onto a Whatman Nytran SPC (super positive charge) nylon membrane. Place a stack of paper towel and 4–5 pieces of Whatman paper in the size of nylon membrane in a tray with 10× SSC soaked nylon membrane.
23. Put the gel on the nylon membrane. Soak two larger pieces of Whatman paper with the transfer buffer and place at the top of the pile. Leave it overnight for the transfer by putting a weight of 400 g on top of the pile.
24. Bake the membrane at 80 °C for 30 min and expose to UV radiation (or by UV cross linker) for 5 min.
25. For preparation of probes, use DIG labeled gene-specific probe to hybridize with plasmid and genomic DNA and leave overnight.

    Preparation of DIG labeled gene-specific probe: Amplify CaMV35S region of the plasmid with PCR and run on agarose gel electrophoresis. Visualize the desired bands on UV transilluminator, cut the PCR product from the gel with a blade, transfer it into a microcentrifuge tube, extract from the gel via a gel extraction kit and labeled with DIG High prime DNA labeling and detection kit.
26. Prehybridize the membrane with prehybridization buffer at 60 °C for 30 min.
27. Take 20 μl probe (If the volume is less, make up to this volume with TE or sterile double distilled water). Denature the probe at 95 °C for 3 min and put on ice for 5 min.
28. Add probe to prehybridization buffer and mix gently and hybridize overnight at 60 °C in hybridization oven.
29. Wash the blot first with low stringency buffer I and then with low stringency buffer II both for 15 min at 65 °C. Then, expose the blot to Kodak BioMax MS film for 2–24 h for autoradiography.

3.4 Agrobacterium-Mediated Transformation of Brachypodium

3.4.1 Preparation of Plant Material and Wounding of Tissues Prior To Transformation

Six-week-old embryogenic calluses induced from mature embryos of *Brachypodium distachyon* lines BdTR4, BdTR6, BdTR13, and Bd21 were used. Wounding of tissues is important for *Agrobacterium*-mediated transformation since it increases the rate of infection and thus the efficiency of transformation. Therefore, we used two different of methods for wounding of callus tissues. One set of calluses were wounded by classical method with a sterile blade (referred to as "macro-wounding"), and microprojectile bombardment-mediated wounding (referred to as "micro-wounding") was used for the other set. Using blade (No. 11), small damages were given to the tissues with its pointed tip. Using the biolistic device, the other set of tissues were bombarded under previously optimized pressure and distances at 650 psi-9 cm with gold particles of 1.0 μm in size and 1100 psi −6 cm with gold particles of 1.6 μm in size. Gold particles were prepared according to the protocol by eliminating the DNA coating step. These two methods were compared to see the difference in transformation efficiency due to wounding type. Micro-wounding of the tissues by microprojectile bombardment increased the transient transgene expression of the infected tissues. On the other hand, bombardment of the callus tissues several times (two or three times) had a negative effect on the transgene expression.

3.4.2 Growth of Agrobacterium tumefaciens Strains

1. Grow AGL1 in MGL medium with addition of antibiotics, kanamycin 100 mg/l and carbenicillin 200 mg/l (*see* **Note 10**).
2. Grow EHA105 and LBA4404 in YEB media with supplementation of following combinations of antibiotics, respectively: kanamycin 100 mg/l and rifampicin 20 mg/l; kanamycin 100 mg/l and streptomycin 100 mg/l.
3. Streak plate the culture from glycerol stocks onto agar plates containing the proper antibiotics and the specified media to get single colonies by incubating at 28 °C. Check that the temperature does not exceed this limit since it could lead to plasmid elimination.
4. Inoculate a single colony into 10 ml liquid medium with antibiotics and incubate at 28 °C at 250 rpm in a shaking incubator. Cultures should be grown in dark by maintaining at least 20% space in the flasks for aeration.
5. Keep the cultures in the incubator more than 24 h and up to 48 h for the activation of bacterial cells (*see* **Note 11**).

3.4.3 Induction of vir Genes

1. Grow 1 ml all bacterial cultures in 50 ml media. Inoculate EHA105 and LBA4404 bacterial culture into YEB-MES medium and AGL1 into MGL medium supplemented with specified antibiotics and 20 μM of acetosyringone as final

concentration. Acetosyringone plays an important role as phenolic compound in the induction of virulence.

2. Measure optical density of the bacterial suspension at $\lambda = 600$ nm and when $O.D._{600}$ is between 0.6 and 0.8, collect the cells by centrifugation at 1400 × *g* for 15 min at 4 °C.
3. Resuspend the cells in MMA medium containing 200 μM acetosyringone and incubate at 22 °C for 1 h or up to 2 h in dark.

3.4.4 Inoculation and Cocultivation

Inoculation and cocultivation steps were optimized according to the procedures applied in the grasses [5, 7, 8, 16, 20, 21].

1. Suspend the wounded callus tissues in bacterial suspension for 15–20 min at 22 °C.
2. Place the calluses with the bacterial cells into vacuum infiltrator for 45–60 min under 50 mmHg pressure to increase the transformation efficiency.
3. Put the calluses on sterile blotting papers to remove the excess bacteria on the plants and cocultivate in MMD medium for 2–3 days at dark under stated tissue culture conditions.
4. After cocultivation period, rinse the tissues 2–3 times first with sterile distilled water before antibiotic washing step.
5. Wash the calluses then with MMA medium containing 500 mg/l cefotaxime for 40 min up to 2 h for the elimination of excess *Agrobacterium*.
6. Place the calluses on sterile blotting papers to remove excess liquid.
7. Transfer the calluses to callus induction medium (CIM), each genotype to its specified medium and keep at least 3–4 days in dark at 25 °C ± 1 to allow the expression of transformed genes (*see* **Note 12**).

3.4.5 Confirmation of Transient Gene Expression, Selection and Regeneration of the Transformed Calluses

1. Carry out transient GUS expression as outlined in Subheading 3.3.5 and use half the number of calluses in each treatment for the assay.
2. Transfer the other half of the calluses that are not subjected to the assay to specified CIM containing 250 mg/l cefotaxime for bacterial inhibition. Supplement the medium also with 3 mg/l PPT for AGL1 transformation or 50 mg/l kanamycin for EHA105 and LBA4404 transformation in order to select the transformed tissues. Culture the calluses for further 2–3 weeks.
3. Carry out selection in magenta boxes with RM containing 250 mg/l cefotaxime, 4 mg/l PPT for AGL1 transformation or 100 mg/l kanamycin for EHA105 and LBA4404 transformation and culture under a 16/8 light/dark photoperiod at

25 ± 1 °C for 1.5–2 months. Subculture to fresh media every 2–3 weeks.

4. Eliminate the necrotic tissues and put the resistant plantlets in RM without any selection. Keep them in magenta boxes under same environmental conditions until the transgenic plants are fully developed. Subculture to fresh media every 2–3 weeks.
5. Transfer the transgenic plants to soil according to acclimatization procedure described in Subheading 3.3.6.

4 Notes

1. Adjust the pH of the tissue culture media using 1 N NaOH and 1 N HCl. During adjustment, all ingredients are included except plant agar. Weigh plant agar and put it into the bottle separately, then pour the pH-adjusted solution into the bottle and autoclave. This will prevent the plant agar from sticking to the probe of the pH meter.
2. The heat-sensitive components that are indicated in the protocol, such as antibiotics, must be filter-sterilized and added to the autoclaved media after cooling down to 50 °C.
3. In order to obtain the explants (stem, root tip, and leaf tissues) rather than mature embryos, seeds are surface-sterilized according to the protocol and regenerated on RM media under specified environmental conditions. The explants are excised after development of the plantlets and placed onto the media.
4. For embryo culture whether using immature or mature embryos, embryos should be excised when used as explant source in tissue culture studies. However, we found an easier way of culturing embryo explants. Whole seeds are exposed to callus induction media and after initiation of callus formation endosperm part of the *Brachypodium* seeds are removed easily from the culture.
5. The embryos may form both fleshy translucent (nonembryogenic) and yellow-hard calluses (embryogenic). We prefer the embryogenic callus tissues since they have higher regeneration capacity. However, if you obtain less embryogenic callus tissue, nonembryogenic tissues can also be used. In addition, as the biomass of the formed callus increases, they should be placed into separate petri plates so that they may have enough area around to take up the nutrients. After subculturing twice, there should be 5–6 callus tissues/petri plate in every plate.
6. About choice of bombardment parameters: Bombardment parameters (size of the gold particle, vacuum pressure,

bombardment pressure, and distance) should be optimized according to the type of species, genotype and explant used.

7. The addition of DNA and chemicals onto gold suspension should be in this order: DNA, $CaCl_2$, and spermidine, since they each have different roles for binding of DNA onto gold particles. During this process, continuous vortexing is important to prevent the settlement of gold particles in gold suspension to bottom of the centrifuge tube.
8. Both shoot and root development of plantlets is achieved using the same regeneration medium (RM). Regeneration is started at the time of plant selection. Antibiotic concentration should be increased gradually since application of high concentration antibiotic at once may damage all cells. However, the time and concentration of application of plant selection agent is important. If you apply the selection with little increases in concentration and prolonged time, the tissues might become resistant to the selection agent and you might have escapes. The non-transformed tissues will regenerate and grow in the selection media.
9. After acclimatization is maintained, the plants can be transferred to green house. You should try to obtain T_1, T_2, T_3, ... transgenic plants and check the stable transformation in the next generations of transgenic lines.
10. *Agrobacterium tumefaciens* strain AGL1 can also be grown in YEB medium, however, growth on MGL medium increases transformation efficiency.
11. *Agrobacterium* growth from glycerol stock takes time at least 24 h and up to 3 days at initial culturing. After the initial growth, it is easier and takes less time to grow during subculturing.
12. Since the supervirulent strains have high capacity to infect, *Agrobacterium* contamination could be observed and washing steps with antibiotic is repeated before transfer of these calluses to new media.

Acknowledgments

This work was supported by TÜBA-GEBİP and Sabancı University.

References

1. Draper J, Mur LAJ, Jenkins G et al (2001) *Brachypodium distachyon*. A new model system for functional genomics in grasses. Plant Physiol 127:1539–1555
2. Ozdemir BS, Hernandez P, Filiz E, Budak H (2008) *Brachypodium* genomics. Int J Plant Genomics 2008:Article ID 536104. doi:10.1155/2008/536104

3. International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768
4. Bablak P, Draper J, Davey MR, Lynch PT (1995) Plant regeneration and micropropagation of *Brachypodium distachyon*. Plant Cell Tissue Organ Cult 42:97–107
5. Vogel JP, Garvin DF, Leong OM, Hayden DM (2006) *Agrobacterium*-mediated transformation and inbred line development in the model grass *Brachypodium distachyon*. Plant Cell Tissue Org Cult 84:199–211
6. Christiansen P, Didion T, Andersen C, Folling M, Nielsen K (2005) A rapid and efficient transformation protocol for the grass *Brachypodium distachyon*. Plant Cell Rep 23:751–758
7. Pacurar DI, Thordal-Christensen H, Nielsen KK, Lenk I (2007) A high-throughput *Agrobacterium*-mediated transformation system for the grass model species *Brachypodium distachyon* L. Transgenic Res 17:965–975
8. Vogel J, Hill T (2007) High-efficiency *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21-3. Plant Cell Rep 27:471–478
9. Vain P, Worland B, Thole V, McKenzie N, Alves SC, Opanowicz M, Fish LJ, Bevan MW, Snape JW (2008) *Agrobacterium*-mediated transformation of the temperate grass *Brachypodium distachyon* (genotype Bd21) for T-DNA insertional mutagenesis. Plant Biotechnol J 6(3):236–245
10. Alves SC, Worland B, Thole V, Snape JW, Bevan MW, Vain P (2009) A protocol for *Agrobacterium*-mediated transformation of *Brachypodium distachyon* community standard line Bd21. Nat Protoc 4(5):638–649
11. Lee MB, Jeon WB, Kim DY, Bold O, Hong MJ, Lee YJ, Park JH, Seo YW (2011) *Agrobacterium*-mediated transformation of *Brachypodium distachyon* inbred line Bd21 with two binary vectors containing hygromycin resistance and GUS reporter genes. J Crop Sci Biotechnol 14(4):233–238
12. Fursova O, Pogorelko G, Zabotina OA (2012) An efficient method for transient gene expression in monocots applied to modify the *Brachypodium distachyon* cell wall. Ann Bot. doi:10.1093/aob/mcs103
13. Trabucco GM, Matos DA, Lee SJ, Saathoff AJ, Priest HD, Mockler TC, Sarath G, Hazen SP (2013) Functional characterization of cinnamyl alcohol dehydrogenase and caffeic acid *O-methyltransferase* in *Brachypodium distachyon*. BMC Biotechnol 13(1):61
14. Filiz E, Ozdemir BS, Budak F, Vogel JP, Tuna M, Budak H (2009) Molecular, morphological and cytological analysis of diverse *Brachypodium distachyon* inbred lines. Genome 52 (10):876–890
15. Nadolska-Orczyk A, Orczyk W, Przetakiewiez A (2000) *Agrobacterium*-mediated transformation of cereals- from technique development to its application. Acta Physiol Plant 22:77–88
16. Jones HD, Doherty A, Wu H (2005) Review of methodologies and a protocol for the *Agrobacterium*-mediated transformation of wheat. Plant Methods 1(5):1–9
17. Murashige T, Skoog F (1962) A revised medium for rapid growth bioassays with tobacco tissue cultures. Physiol Plant 15:473–497
18. Jefferson RA (1987) Assaying chimeric genes in plants: The GUS gene fusion system. Plant Mol Biol Rep 5:387–405
19. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull 19:11–15
20. Li L, Li R, Fei S, Qu R (2005) *Agrobacterium*-mediated transformation of common bermudagrass (*Cynodon dactylon*). Plant Cell Tissue Org Cult 83:223–229
21. Luo H, Hu Q, Nelson K, Longo C et al (2004) *Agrobacterium tumefaciens*-mediated creeping bentgrass (*Agrostis stolonifera* L.) transformation using phosphinothricin selection results in a high frequency of single-copy transgene integration. Plant Cell Rep 22:645–652

INDEX

Gaurav Sablok et al. (eds.), *Brachypodium Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1667, DOI 10.1007/978-1-4939-7278-4, © Springer Science+Business Media LLC 2018

O

P

R

S

T

W

X