Springer Protocols

Methods in Molecular Biology 513

Plant Genomics

Methods and Protocols

Edited by

Daryl J. Somers
Peter Langridge
J. P. Gustafson

Humana Press

METHODS IN MOLECULAR BIOLOGY™

Plant Genomics

Methods and Protocols

Edited by

Daryl J. Somers*, Peter Langridge†, and J. Perry Gustafson‡

*Molecular Breeding and Biotechnology, Vineland Research and Innovation Centre, Vineland Station, Ontario, Canada**
Australian Centre for Plant Functional Genomics, University of Adelaide, Glen Osmond, Australia†
Division of Plant Sciences, University of Missouri, Columbia, MO, USA‡

Humana Press

Editors
Daryl J. Somers
Molecular Breeding and Biotechnology
Vineland Research and Innovation Centre
Vineland Station, Ontario
Canada

J. Perry Gustafson
Division of Plant Sciences
University of Missouri
Columbia, MO, USA

Peter Langridge
Australian Centre for Plant Functional Genomics
University of Adelaide
Glen Osmond, Australia

ISBN: 978-1-58829-997-0 e-ISBN: 978-1-59745-427-8
ISSN: 1064-3745 e-ISSN: 1940-6029
DOI: 10.1007/978-1-59745-427-8

Library of Congress Control Number: 2008940985

Printed on acid-free paper

springer.com

METHODS IN MOLECULAR BIOLOGY™

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

For other titles published in this series, go to
www.springer.com/series/7651

Preface

This volume is divided into chapters which consider the primary issues and methodologies surrounding plant genomics research. Plant genomics is largely concerned with associating functional genes or gene mutations with phenotype. Therefore, chapters are included that cover the areas of gene discovery and functional analysis of genes. Further chapters focus on the primary tools and sub-disciplines of genetic mapping, mRNA, protein and metabolite profiling. Methods are included that explore gene functional analysis via transformation, mutation, protein function and gene expression. The volume includes chapters on data management which consider the expansion of plant genomics databases and bioinformatics analysis tools. The volume is concluded with chapters aimed at discussing the application and deployment of molecular plant breeding technology from the use of markers in breeding, development of genetically modified plants/crop species, analysis of existing populations for novel alleles and gene/trait associations and genome sequencing.

Contributors

Olin D. Anderson • *Western Regional Research Center, Albany, CA, USA*

Nikki Appleby • *Australian Centre for Plant Functional Genomics, Institute for Molecular Biosciences and School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia*

Jacqueline Batley • *Australian Centre for Plant Functional Genomics, Institute for Molecular Biosciences and School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia*

C. Robin Buell • *The Institute for Genomic Research, Rockville, MD, USA*

Rachel A. Burton • *Australian Centre for Plant Functional Genomics, School of Agriculture, Food and Wine, University of Adelaide, Glen Osmond, SA, Australia*

Natalia V. Bykova • *Agriculture and Agri-food Canada, Cereal Research Centre, Winnipeg, MB, Canada*

Kevin Childs • *The Institute for Genomic Research, Rockville, MD, USA*

Molly N. Cline • *Monsanto, St. Louis, MO, USA*

Paul Derbyshire • *John Innes Centre, Norwich, UK*

Angela Doherty • *CPI Division, Rothamsted Research, Harpenden, Hertfordshire, UK*

Liam Dolan • *John Innes Centre, Norwich, UK*

John H. Doonan • *John Innes Centre, Norwich, UK*

Sinéad Drea • *Department of Molecular, Cell and Developmental Biology, Yale University, New Haven, CT, USA*

Chris Duran • *Australian Centre for Plant Functional Genomics, Institute for Molecular Biosciences and School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia*

Arnis Druka • *Scottish Crop Research Institute, Invergowrie, Dundee, Scotland, UK*

David Edwards • *Australian Centre for Plant Functional Genomics, Institute for Molecular Biosciences and School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia*

Naser Farrokhi • *Department of Biological Sciences, California State University, Long Beach, CA, USA*

Geoffrey B. Fincher • *Australian Centre for Plant Functional Genomics, School of Agriculture, Food and Wine, University of Adelaide, Glen Osmond, SA, Australia*

David A. Fischhoff • *Monsanto, St. Louis, MO, USA*

Richard Flavell • *Ceres, Inc., Thousand Oaks, CA, USA*

Andreas Graner • *Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany*

Sean Grimmond • *Institute for Molecular Biosciences, University of Queensland, Brisbane, Australia*

J. Perry Gustafson • *Division of Plant Sciences, University of Missouri, Columbia, MO, USA*
Dave A. Hoisington • *International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India*
Maria Hrmova • *Australian Centre for Plant Functional Genomics, School of Agriculture, Food and Wine, University of Adelaide, Glen Osmond, SA, Australia*
Michael Imelfort • *Australian Centre for Plant Functional Genomics, School of Land, Crop and Food Sciences, University of Queensland, Brisbane, Australia*
Huw D. Jones • *CPI Division, Rothamsted Research, Harpenden, Hertfordshire, UK*
Michael J. Kearsey • *School of Biosciences, University of Birmingham, Birmingham, UK*
Rachil Koumproglou • *John Innes Centre, Norwich, UK*
Peter Langridge • *Australian Centre for Plant Functional Genomics, University of Adelaide, Glen Osmond, Australia*
Gerard R. Lazo • *Western Regional Research Center, Albany, CA, USA*
David E. Matthews • *Department of Plant Breeding, Cornell University, Ithaca, NY, USA*
Spurthi N. Nayak • *International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India*
Eviatar Nevo • *Institute of Evolution and the International Graduate Center of Evolution, University of Haifa, Mount Carmel, Haifa, Israel*
Shu Ouyang • *The Institute for Genomic Research, Rockville, MD, USA*
Isobel A. P. Parkin • *Agriculture and Agri-Food, Saskatoon Research Centre, Saskatoon, SK, Canada*
Elena Potokina • *School of Biosciences, University of Birmingham, Birmingham, UK*
Christof Rampitsch • *Agriculture and Agri-food Canada, Cereal Research Centre, Winnipeg, MB, Canada*
Stephen J. Robinson • *Agriculture and Agri-Food, Saskatoon Research Centre, Saskatoon, SK, Canada*
Bu-Jun Shi • *Australian Centre for Plant Functional Genomics, University of Adelaide, Glen Osmond, Australia*
Peter Shaw • *John Innes Centre, Norwich, UK*
Caroline A. Sparks • *CPI Division, Rothamsted Research, Harpenden, Hertfordshire, UK*
Françoise Thibaud-Nissen • *The Institute for Genomic Research, Rockville, MD, USA*
Rajeev K. Varshney • *International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India*
Wei Zhu • *The Institute for Genomic Research, Rockville, MD, USA*

Contents

Chapter 1

Role of Model Plant Species

Richard Flavell

Summary

The use of model or reference species has played a major role in furthering detailed understanding of mechanisms and processes in the plant kingdom over the past 25 years. Species which have been adopted as models for dicotyledons and monocotyledons include *arabidopsis* and rice and more recently *brachypodium*, Such models are diploids, have few and small chromosomes, well developed genetics, rapid life cycles, are easily transformed and have extensive sets of technical resources and databases curated by international resource centres. The study of crop genomics today is deeply rooted in earlier studies on model species. Genomes of model species share reasonable genetic synteny with key crop plants which facilitates the discovery of genes and association of genes with phenotypes. While some mechanisms and processes are conserved across the plant kingdom and so can be revealed by studes on any model species, others have diverged during evolution and so are revealed by studying only a closely related model species. Examples of processes that are conserved across the plant kingom and others that have diverged and therefore need to be understood by studying a more closely related model species are described.

Key words: Genomes, Synteny, Comparative genomics, Genome sequence.

1. Introduction

Evolutionary and comparative genetics between plant species has validated the use of one species as a model for another, for the purpose of understanding plant biology. The process of deliberately selecting "model" species over the last two decades, suitable for amassing information rapidly and cheaply by thousands of scientists, has provided a revolution in our understanding of plants. The complete genome sequences and gene–trait associations revealed for these species has provided enormous insight into all plant species, their chromosomes, genes, pathways, evolution and hence relationships to one another and has provided an early framework

Daryl J. Somers et al. (eds.), *Methods in Molecular Biology, Plant Genomics, vol. 513*

DOI: 10.1007/978-1-59745-427-8_1

for understanding the genetic and molecular diversity in plants and plant processes. Yet, it is only a beginning because of the immense diversity across the plant kingdom. Because of this diversity, the concept of one or a few species being "models" suitable for all species is flawed. The major challenges are therefore (*1*) to evaluate the current framework gained from the relatively few "model" species, (*2*) to use the framework to understand many species, recognizing both the strengths and weaknesses of the framework for comparative biology and (*3*) to extend the framework by studying additional, specially selected, species based on plant phylogeny.

While at any one-time model species are useful for providing predictions relevant to other members of the plant kingdom, they leave, of course, the need to test the predictions for any particular species, for example, the crop species that provide our food, feed, fiber and energy. However, the framework of understanding gained from selected "model" species is a wonderful starting point to evaluate any species in detail with speed and insight.

2. History

It was during the 1980s when plant scientists worldwide were studying processes and traits in a very large range of plant species, especially economically important species, that it became accepted both in the scientific community and the funding agencies, in the EU and USA particularly, that much more benefit could be gained by focusing on one or two species as models for crops and processes across the plant kingdom. It was controversial because the models being touted were not economically important crops and it meant fewer funds for the favourite and important crops such as maize, tomato, wheat and barley about which a lot of information was being gathered. Yet, it had become obvious that having a large number of scientists studying *Escherichia coli*, yeast, *Drosophila* and *Homo sapiens* produced so much more detailed and understood information that knowledge of plants, important as they are, was being left behind. In consequence, the most talented minds were not being attracted to plant biology on the same scale as to model organisms. It had also become obvious that it was going to be possible to sequence whole plant genomes to unleash the power of genomics and so debates arose as to which genome would be sequenced and how the results would be used. The molecular genetics approaches of the models mentioned above were the most appealing especially also because plant breeding is based on genetics and genomics. Thus, the vision was adopted to learn the sequences of all the genes in some model plant and determine their function via mutational genetics and reverse genetics.

An ideal model needs to be able to be studied to give rise to relevant information more quickly and cheaply than studying other species *(1, 2)*. Some of the key features of an initial model are shown in **Table 1**. Speed, cost and convenience are key features. They drive scientists and funding agencies, especially in this day and age of the competitive environments in which there is a need to demonstrate substantial progress in a very short time. With these features being fulfilled in a model, it is impossible for an equivalent number of experiments to be done on more cumbersome species.

In the 1980s, fulfilling the vision appeared possible only with a diploid species that had a small genome, a rapid life cycle and that could easily be transformed with novel genes. Many other factors also held a place in the debate, including how easy it was to grow the plant in a small environment. These are the reasons why *Arabidopsis* became the leading contender around the world *(3–6)* after some debate about *Petunia* and some other species. Friedrich Laibach had studied *Arabidopsis* from the early 1900s, and Erna Rheinholz in the early 1940s, but it was Glass *(7)*, Redei *(8)* and Koornneef *(9)* who opened up mutational genetics in the species.

While the genomics-based approaches were being developed for *Arabidopsis*, mainly in USA and Europe, rice genomics was being driven, especially in Asia and USA, by the importance of rice as a crop and the fact that its genome is also small and strains of rice are easily transformable. The "full" japonica genome sequence was published in 2002 *(10, 11)* with several updates being published subsequently from the international sequencing consortium including telomere repeats (http://rgp.dna.affrc.go.jp) and the sequence of centromeres *(12)*.

Table 1
Preferred attributes of a model crop species

Attributes
Small genome
Rapid life cycle
Easily transformed
Diploid genetics with few chromosome/gene duplications
Well positioned in plant phylogeny
Small stature for growth in small space
Large number of seeds produced
Convenient for discovery of gene–trait linkages at low cost, high speed

Arabidopsis, classified within the eudicots lineage of flowering plants, inevitably has major limitations as both a model and a framework reference for monocots that occur in the other major lineage of flowering plants (**Fig. 1**). That is why rice plays such an important role for understanding monocots and monocot genomes, and complements *Arabidopsis* for studying angiosperms in general. While experiments with rice are not as fast and as cheap as *Arabidopsis*, the large volume of work being done in Asia has resulted in a lot being achieved at a fast pace. Much of the thinking behind the experimental approaches was learnt from *Arabidopsis*, which, in turn, was modelled after yeast, *Drosophila* etc.

While the genomics of other species has been initiated, they have intrinsic difficulties that prevent such rapid progress in genetics, gene–trait linkages and developmental biology compared with rice and *Arabidopsis*. Nevertheless, poplar has been adopted as a model for trees since some strains of it are readily transformable and the US Department of Energy's Joint Genome Institute (JGI, www.jgi.doe.gov) has completed the sequence of its genome *(13–15)*. The sorghum genome has been recently sequenced by the JGI and that of corn is well advanced, as is

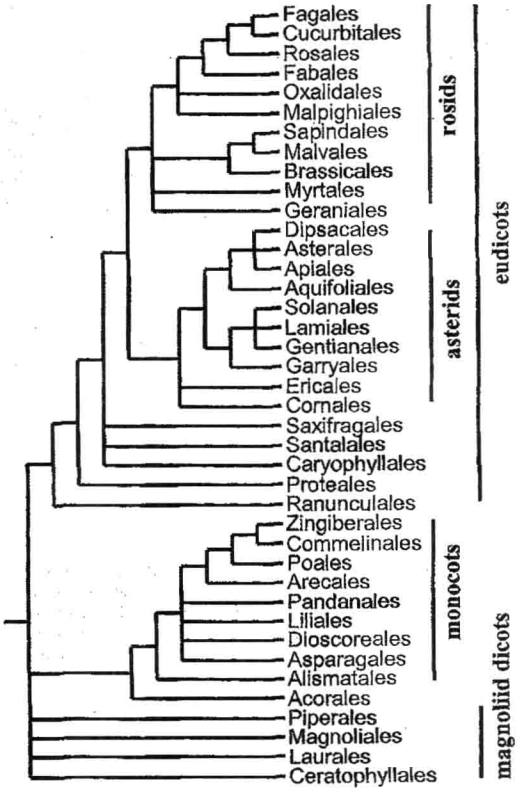

Fig. 1. Angiosperm phylogeny modified from Angiosperm Phylogeny Group *(65, 66)*. *Arabidopsis* is in *Brassicales* of the rosids, and rice is in *Poales* of the monocots.

that of Medicago which can serve as a model for certain legumes. The genome of *Brachypodium distachyon* is also being sequenced. This species, with its small genome and relative ease of transformation, has been adopted recently as a model for temperate C3 monocot grasses that will hopefully provide information particularly relevant to wheat, barley and other grasses *(16*, www.brachypodium.org).

The success of *Arabidopsis* as the leading model species and its value can be inferred from the number of publications and the databases devoted to the species since 1985. In those days just a few dozen papers per year were published on *Arabidopsis.* In 2006, there were more than 2,200 in peer-reviewed journals *(17).* The *Arabidopsis* Information Resource (TAIR, *18)* reports that there are now ~16,000 *Arabidopsis* researchers in about 6,200 laboratories worldwide. They are linked together under the auspices of "The Multinational Coordinated *Arabidopsis thaliana* Functional Genomics Project" (MCAtFGP) that publishes an update each year. These statistics mean that *Arabidopsis* has attracted much competitive grant money and people to devote their research careers to the study of the model plant. The initiative has had an enormous impact on plant biology. Spending the same time and amount of money could not have led to anything like our current understanding of plant biology had we continued in the same way as prior to the early 1980s. The 2007 report of the MCAtFGP makes the case as follows: "Research on Arabidopsis has provided most of the breakthroughs made in plant science over the last ten years and, given the continuing rapid progress, will drive the major discoveries in plant science for the next ten years. The resources and expertise are available to meet the goal of discovering a function for all the Arabidopsis genes of major significance within a reasonable timeframe. Given a high level of continuing support over several decades the ultimate goal of obtaining a working understanding of how a flowering plant functions down to a molecular level is within sight. Such a working model would be of incalculable benefit to future generations of scientists, farmers, environmentalists and society at large." The major claim that "Arabidopsis has provided most of the breakthroughs over the past ten years" is a very bold one but accurate overall, illustrating the impact of this model on the molecular genetics of plants.

3. Genomics, Tools and Databases for *Arabidopsis* and Rice

The selection of *Arabidopsis* and rice as the principal models with which to develop, rapidly and cheaply, understanding of plant biology went hand in hand with the completion of full genome

sequences (http://plantgdb.org/AtGDB,*19)*, collections of full length cDNAs *(18, 20)*, descriptions of expressed genes via deep EST sequencing, development of the use of microarrays and deep signature sequencing (www.dbi.udel.edu) to study gene expression patterns in different organs and growth conditions, the production of stocks with T-DNA mutations in "every" gene, stocks with transgenes inserted, recombinant inbred lines and mapping populations, molecular markers for quantitative trait loci (QTL) mapping and much more. These are detailed on The Arabidopsis Information Resource (TAIR) website for *Arabidopsis* and on The Rice Genome Resource Center website for rice http://www.rgrc.dna.affrc.go.jp/ and are described in part in other chapters of this book (*see* also *21, 22)*. The physical resources for *Arabidopsis* and rice have been deposited in stock centres to facilitate curation, QC and access for all (http://arabidopsis.info;www.biosci.ohio-state.edu/pcmb/facilities/abrchome.htm;http://www.rgrc.dna.affrc.go.jp/). Similarly databases describing the compendium of genomics information have been established from the beginning (*see* TAIR). These open access tools and databases have been of extraordinary value to drive forward the development and use of these species as models. For *Arabidopsis*, they were associated with goals set by the scientific community and the US National Science Foundation to, for example, find the function of every gene, and now micro RNA *(23)*, by 2010 *(24)*. The forward-looking research emphases are on the networks formed by the physical, genetic, metabolic and regulatory interactions between genes, proteins and metabolites.

The very large number of experiments assessing the levels of expression of *Arabidopsis* genes under many different conditions in different organs (*see* TAIR) is a wonderful resource for addressing the functions of genes, networks and genes that are co-regulated. These databases are also useful for selecting promoters with specific expressions patterns.

The complete genome sequences of different accessions of *Arabidopsis* and rice are also being determined to better understand mutational events and variation in populations and, in association with QTL mapping, to link variation in genes with traits. Over 250,000 high quality single nucleotide polymorphisms (SNPs) are available from sequencing several *Arabidopsis* accessions (*see* TAIR). Recently, *Arabidopsis* genomics research has led the way in describing a global view on methylation patterns using high resolution tiling microarrays *(25)* to add to the fast growing field of epigenetics.

With all this data there is special emphasis on data storage, analysis and visualization. This requires the formation of user-friendly databases and development of annotations that are adopted across species. Descriptions of genes and processes in different species must be harmonized to enable comparisons to

be made with accuracy. This has not historically occurred in gene description terms. *Arabidopsis* descriptors based on chromosome location provide unambiguous reference points, but these are meaningless for across-species comparisons. However, the Gene Ontology terminology is an attempt to provide such terms and is being developed for plants *(26*, www.geneontology.org).

The combined use of genetic variation and phenotypic screens has been developed in a huge number of ways to gain a primary understanding of gene–trait relationships. Three sorts of approaches have been adopted. First, and the most widely used has been to screen large populations of mutants with T-DNA (*see* TAIR) or transposon *(27)* insertions to find the variant which has the desired phenotypic change and then to sequence around the T-DNA/transposon insert in the selected plant to find the gene into which it has inserted (e.g., *21, 28, 29)*. While the approach has been very successful, the fact that mutations often occur during transformation at sites other than where the T-DNA is inserted, and that multiple T-DNAs are frequently inserted means that tests to check the complete linkage between the T-DNA/transposon and phenotype must be carried out. Alternatively, multiple T-DNA/transposon insertions at the same locus, causing the same phenotype, can be obtained to establish the gene–trait association. Failure of studies with T-DNA/transposon insertion mutants to identify a phenotypic change can be due to (1) the screens deployed not being appropriate or (2) that the mutated gene is duplicated in the genome and so mutations in all members of the gene family would be required to see the phenotypic effect. The second approach has been so-called "activation tagging" *(28, 30)*, where T-DNAs carrying a strong enhancer of expression are inserted into plant genomes at a very large number of locations, with the assumption that when an enhancer inserts close to a gene the gene will be activated and phenotypic changes will give a gene–trait association for that gene. Populations carrying the enhancers are screened, plants with desired phenotypes selected, the genomic location of the T-DNA(s) determined and nearby genes examined for altered expression. The genes can then be tested individually for their ability to cause similar phenotypic changes when expressed at higher levels and/or in different cells. The third approach, which has been widely adopted by many, includes the companies Ceres (www.ceres.net) *(2)*, Monsanto (www.monsanto.com) Mendel (www.mendelbio.com) and Icoria (www.icoria.com) (now Monsanto). The third approach has also been adopted by Crop Design (now BASF) for rice. These companies have operated high throughput strategies, exploiting the ease of transformation of *Arabidopsis*, to mis-express large numbers of transgenes under the control of very active promoters and then to screen the resulting plants for changes in defined traits. Genetic variation emanating from changes in the level of

expression might be equivalent to that frequently occurring in natural populations as well as in breeding (crop improvement) populations. Where the mis-expressed gene is from another species then the protein sequence is different from that in *Arabidopsis* and so the effects of this variation can also be scored. Failure of mis-expression to cause a detectable phenotype can be because (1) the amount of RNA and protein being expressed is not affecting the networks that link expression of the gene with the manifested trait, (2) the screen is not examining the relevant trait, or (3) changes in the levels of expression of multiple genes are required to create a phenotypic change. In this situation, no conclusions about the role of the gene in a trait can be drawn. With this approach there is the possibility that the phenotypes are due to over-expression of homologous gene silencing due to the formation of double stranded RNA from the transgene insert or cluster of inserts. Typically not all transformants show the same phenotype and this opens up the possibility of multiple mechanisms for causing a change in phenotype.

Tens of thousands of full length cDNAs as well as genomic DNAs have been put through this regime and morphological phenotypes, including flowering time, scored visibly and in over 20 screens covering a wide range of stresses, including drought, salt, heat, cold tolerance, low nitrogen, high and low light, traits very important in applied plant breeding. These screens have taken advantage of the small size of *Arabidopsis* and the ability to evaluate the plants in growth rooms, greenhouse, in soil and on defined media in petri dishes. They could not be done easily or cheaply on this scale with larger plants. This illustrates the very special advantage of *Arabidopsis* for such studies. The experiments developed on this scale also required a very efficient pipeline of gene cloning, plant transformation, seed collection and screening coupled with efficient sample tracking and data collection.

All of these approaches have led to knowledge of hundreds or thousands of gene–trait linkages, some by loss of gene function and others by activation of gene function. When a gene–trait linkage has been found it can be checked by evaluating independent transgenic events and showing strict inheritance of the trait with the transgene over generations.

These gene–trait linkages are clearly defined by the specific genetic background of the accession of the model species used. How useful is the genetic background of such a model species, selected for the speed and cost of doing the experiments, for predicting gene–trait linkages in other species that have diverged significantly from the models during evolution? This is a key question because the answer will determine the extent to which the use of models will be of direct utility to applied plant breeding.