R.J. Schnell · P.M. Priyadarshan *Editors*

# Genomics of Tree Crops

Springer

Genomics of Tree Crops

R.J. Schnell · P.M. Priyadarshan
Editors

# Genomics of Tree Crops

Springer

*Editors*
R.J. Schnell
USDA - ARS
Miami, FL, USA

P.M. Priyadarshan
Rubber Research Institute of India
Central Experiment Station
Ranni, Kerala, India

ISBN 978-1-4614-0919-9 ISBN 978-1-4614-0920-5 (eBook)
DOI 10.1007/978-1-4614-0920-5
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012941613



Printed on acid-free paper

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

# Preface

Editing a book on Genomics is a difficult task since this branch of science is changing so quickly. However, it is necessary to compile the advancements in this area for the use of students and new scientists interested in tree species. Many technological innovations have occurred in a short time span and understanding these techniques and how they are used is essential to move the science of plant genetics forward. Since the rediscovery of Mendelism during the 1900s, the genetics and breeding of plants has experienced many paradigm shifts such as the discovery of DNA and RNA, unraveling of genes and gene expression, the central dogma, reverse genetics, gene sequencing, molecular genetic markers, polymerase chain reaction (PCR), anti-sense RNA, single nucleotide polymorphisms (SNP), RNA interference (RNAi), epigenetics, and finally functional genomics.

In the last few decades, the pace of biological research has accelerated as we have increased our ability to manipulate genes. In agriculture, this has led to waves of controversy, but in medicine the advances are almost universally applauded. Regardless of one's views on genetic engineering, no one questions that it is changing the science of biology in profound ways.

The completion of the Human Genome Project was reported in the first year of the new millennium, with the full sequence becoming available for research and exploration. The Human Genome Project, initiated in the late 1980s, determined the entire sequence of three billion nucleotides. Major goals of this project were to identify and understand a whole repertoire of human genes. During the next few years, the function of all of the estimated 50,000–100,000 human genes will be identified. Embryonic stem cells were cloned for the first time in 2000, and offer the potential for curing a wide range of ills, from spinal cord injuries to diabetes. Golden rice, a genetically modified crop to which a battery of genes that overcome vitamin A and iron deficiencies have been added, was planted for the first time in Asian fields. Even taxonomy seems to be undergoing a sea change, with molecular phylogenies forcing the redrawing of many family trees, from angiosperms to insects and other arthropods.

Molecular Plant Breeding, the new science that emerged from plant genomics, presents many opportunities: shortening the time it takes to domesticate new crops

from semi-wild plants, tailoring existing crops to meet new requirements such as nutritional enhancement or resistance to climate change, rapidly incorporating valuable traits from wild relatives into established crops, allowing plant breeders to work with highly complex traits, such as hybrid vigor and flowering, and making it feasible to work on research-neglected "orphan" crops.

Conceptually, whole genome sequencing represents an ultimate form of reductionism in molecular biology. The complex processes of life cannot be totally explained by the linear sequence of DNA. In experimental reality, DNA sequencing requires drastic reductions from higher to lower dimension – to destroy the cell and to extract the DNA molecules. We do not question how much information is lost in these procedures, but simply accept the common wisdom that the genome, or the entire set of DNA molecules, contains all the necessary information to make up the cell.

Genome projects have transformed biology in many ways, but the most immediate outcome is the emergence of computational biology, also known as bioinformatics. It is no longer possible to make advances in biology without integration of informatics technologies and experimental technologies. There is a distinction between genome informatics and post-genome informatics here. Genome informatics was born of necessity to cope with the vast amount of data generated by the genome projects. In contrast, post-genome informatics represents a synthesis of biological knowledge from genomic information toward understanding the basic principles of life. Post-genome informatics has to be coupled with systematic experiments from a large range of scientific disciplines to understand and manipulate plant metabolism for human advantage. This understanding is essential for mitigation of a host of environmental interactions, including physical and biological stress, that are likely to have a greater influence with the accelerating climate change occurring around the world.

The genetic enhancements made in annual crop species stand in sharp contrast to those achieved for tree species. Trees are genetically recalcitrant and have long generation times. They require large amounts of land for phenotypic evaluation and have much less capital resources provided for plant improvement programs. As a result, much less is understood and progress in tree genomics is meager. It is this fact that made us undertake the editing of a book on Genomics of Tree Crops. Our goal was to bring out a compilation of the progress made in tree crops. We have chapters on: the state of the art, bioinformatics, functional genomics of flowering time, gene flow, spatial structure and local adaptation, genetic transformation of fruit trees, genomics of tropical and temperate fruit trees, papaya genomics, genomics of *Hevea* rubber, and genomics of palms. These chapters are contributed by experts in their respective areas of specialization.

We thank Springer for agreeing to publish this book.

Miami, FL, USA — R.J. Schnell
Kerala, India — P.M. Priyadarshan

# Contents

# Contributors

**María José Aranzana, Ph.D.** Department of Plant Genetics, IRTA – Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Barcelona, Spain

**Renée S. Arias, Ph.D.** National Peanut Research Laboratory, USDA-ARS, Dawson, GA, USA

**James W. Borrone, MS, Ph.D.** Department of Entomology & Plant Pathology, Oklahoma State University, Stillwater, OK, USA

**Jaroslaw Burczyk** Department of Genetics, Institute of Experimental Biology, Kazimierz Wielki University, Bydgoszcz, Poland

**Igor Chybicki, Ph.D.** Department of Genetics, Institute of Experimental Biology, Kazimierz Wielki University, Bydgoszcz, Poland

**Gregory Downs, B.Sc.** Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada

**Iban Eduardo, Ph.D.** Department of Plant Genetics, IRTA – Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Barcelona, Spain

**Reiner Finkeldey** Büsgen-Institute, Forest Genetics and Forest Tree Breeding, Georg-August University of Göttingen, Göttingen, Lower Saxony, Germany

**Henryk Flachowsky** Institute for Breeding Research on Horticultural and Fruit Crops, Julius Kuehn-Institute, Dresden, Saxony, Germany

**Matthias Fladung, Ph.D.** Institute of Forest Genetics, Johann Heinrich von Thünen Institute, Grosshansdorf, Germany

**Magda-Viola Hanke** Institute for Breeding Research on Horticultural and Fruit Crops, Julius Kuehn-Institute, Dresden, Saxony, Germany

**Hans Hoenicka** Institute of Forest Genetics, Johann Heinrich von Thünen Institute, Grosshansdorf, Germany

**Brian M. Irish, BS, MS, Ph.D.** Tropical Agriculture Research Station, USDA-ARS, Mayaguez, PR, USA

**Maizura Ithnin, Ph.D.** Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, Kajang, Selangor, Malaysia

**Robert R. Krueger, Ph.D.** USDA-ARS National Clonal Germplasm Repository for Citrus & Dates, Riverside, CA, USA

**Konstantin V. Krutovsky, Ph.D.** Department of Ecosystem Science and Management, Texas A&M University, College Station, TX, USA

**David N. Kuhn, Ph.D.** Subtropical Horticulture Research Station, USDA-ARS, Miami, FL, USA

**Richard E. Litz, Ph.D.** Tropical Research & Education Center, University of Florida, Homestead, FL, USA

**Eng-Ti L. Low, Ph.D.** Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, Kajang, Selangor, Malaysia

**Lewis Lukens, Ph.D.** Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada

**Ana Montserrat Martín-Hernández** Department of Plant Genetics, IRTA – Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Barcelona, Spain

**Alan W. Meerow, Ph.D.** USDA-ARS-SHRS, National Germplasm Repository, Miami, FL, USA

**Ray Ming, Ph.D.** Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, USA

**Paul H. Moore, Ph.D.** Hawaii Agriculture Research Center, Kunia, HI, USA

**Leslie C.-L. Ooi, M.Sc.** Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, Kajang, Selangor, Malaysia

**Guillermo Padilla, Ph.D.** Agrobiología y Medio Ambiente, Instituto de Productos Naturales y Agrobiología (IPNA-CSIC), San Cristóbal de La Laguna, Santa Cruz de Tenerife, Spain

**P.M. Priyadarshan, Ph.D.** Rubber Research Institute of India, Central Experiment Station, Ranni, Kerala, India

**Tanja Pyhäjärvi, Ph.D.** Department of Plant Sciences, University of California, Davis, CA, USA

**Juan Jose Robledo-Arnuncio, Ph.D.** Departamento de Ecología y Genética Forestal, Centro de Investigación Forestal (CIFOR) – INIA, Madrid, Spain

**Carlos Romero** Citricultura y Producción Vegetal, Instituto Valenciano de Investigaciones Agrarias, Moncada, Valencia, Spain

**Thakurdas Saha, Ph.D.** Genome Analysis Laboratory, Rubber Research Institute of India, Kottayam, Kerala, India

**Raymond J. Schnell, Ph.D.** National Germplasm Repository for Tropical/ Subtropical Fruit Crops, USDA-ARS, Miami, FL, USA

**Rajinder Singh, Ph.D.** Advanced Biotechnology and Breeding Centre, Malaysian Palm Oil Board, Kajang, Selangor, Malaysia

**Cecile L. Tondo, Ph.D.** Subtropical Horticulture Research Station, USDA-ARS, Miami, FL, USA

**Santiago Vilanova** Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana, Universidad Politécnica de Valencia, Valencia, Spain

**Qingyi Yu, Ph.D.** Texas Agrilife Research Center, Department of Plant Pathology and Microbiology, Texas A&M University, Weslaco, TX, USA

# Chapter 1
# The State of the Art: Molecular Genomics and Marker-Assisted Breeding

**P.M. Priyadarshan and Raymond J. Schnell**

**Abstract** Focus on tree biotechnology reflects the challenges posed by the genetic attributes of trees. The genetic attributes of trees stand in stark contrast to those of domesticated annual crops. Trees typically have long generation times and are wind pollinated with out-crossing mating systems. Traditional tree breeding is a lengthy process that cannot efficiently capture nonadditive genetic variation, primarily because inbred lines would suffer from inbreeding depression. Clonal propagation of elite genotypes allows for the capture of both additive and nonadditive genetic variation, and the addition of transgenes can confer new or enhanced traits. The molecular analysis of plants often focused on the single gene level. But the recent technological advances have changed this paradigm. The way the genes and genetic information are organized within the genome and the methods of collecting and analyzing this information and the determination of their biological functionality are referred to as genomics. Genomic approaches are permeating every aspect of plant biology, and since they rely on DNA-coded information, they expand molecular analyses from a single to a multispecies level. Plant genomics is reversing the previous paradigm of identifying genes behind biological functions and instead focuses on finding biological functions behind genes. It also reduces the gap between phenotype and genotype. This introductory chapter overviews two main sections: first, the current understanding of genomes, their genetic structure at the inter- and intra-species level, and how whole genomes are sequenced; and second, on finding the biological and functional significance of DNA sequence. It is also

P.M. Priyadarshan, Ph.D. (✉)
Rubber Research Institute of India, Central Experiment Station,
Chethackal, Thompikandom, Ranni, Kerala 689676, India
e-mail: rriipriya@gmail.com

R.J. Schnell, Ph.D.
National Germplasm Repository for Tropical/Subtropical Fruit Crops,
USDA-ARS, Miami, FL, USA
e-mail: Ray.Schnell@effem.com

R.J. Schnell and P.M. Priyadarshan (eds.), *Genomics of Tree Crops*,
DOI 10.1007/978-1-4614-0920-5_1, 

worthwhile to note that these technologies, though extensively used in agricultural species, are only used in forest tree species research. Except for some tropical (avocado, mango, and papaya) and temperate (apple, *Prunus*, and *Pyurus*) fruit species, these techniques are not extensively used in other tree crops.

**Keywords** Breeding • DNA • Genetic markers • Genomes • Genetic maps • Reverse genetics • Transcriptional profiling

## Introduction

Focus on tree biotechnology reflects the challenges posed by the genetic attributes of trees and the limitations of extending available biotechnologies developed for agricultural crop species to trees. The genetic attributes of trees stand in stark contrast to those of domesticated annual crops. Trees typically have long generation times and are wind pollinated with out-crossing mating systems. Individual trees are highly heterozygous and thus carry a high genetic load, such that mating between related individuals results in inbreeding. Furthermore, unlike crop species, trees are expected to have minimal population substructure and low linkage disequilibrium (Gonzalez-Martinez et al. 2006; Ingvarsson 2005). A practical consequence of low linkage disequilibrium is that linkage relationships between markers and alleles of genes controlling phenotypic traits are not consistent among individuals, which limits the application of marker-assisted selection and breeding.

Traditional tree breeding is a lengthy process that cannot efficiently capture non-additive genetic variation, primarily because inbred lines would suffer from inbreeding depression. Clonal propagation of elite genotypes allows for the capture of both additive and nonadditive genetic variation, and the addition of transgenes can confer new or enhanced traits. For example, damage from introduced diseases and insects for which there is no natural genetic basis for resistance could be mitigated through introduction of transgenes conferring resistance (Adams et al. 2000). However, research on the strategies and risks of introducing transgenics into natural populations is still in its infancy (DiFazio et al. 2004; van Frankenhuyzen and Beardmore 2004). Political, societal, and regulatory restrictions make the application of transgenics to trees in the near future uncertain (Herrera 2005). The lengthy traditional tree breeding process typically relies on identifying trees with desirable attributes, followed by indirectly evaluating their breeding potential by measuring phenotypic traits in their progeny. Most traits of interest to forest industry are quantitative in nature, can be costly to measure, and occur later in development (e.g., wood quality). To better understand the genetic regulation of quantitative traits and speed up the progeny testing process, research has focused on the ability to detect chromosomal regions carrying favorable alleles controlling quantitative traits, so called quantitative trait loci (QTL). Studies on tree species have demonstrated the feasibility of this approach within pedigrees and have identified quantitative trait loci influencing traits ranging from wood properties to adaptive traits (Jermstad et al. 2003). Marker-assisted selection is an extension of QTL technology, in which progeny with desired

genotypes within a given pedigree are identified using molecular markers linked to favorable QTL alleles. However, QTL and marker-aided selection have limited application outside of pedigreed material. Limitations to QTL and marker-aided selection are exposed when consideration is given to the low linkage disequilibrium and high allelic variation present especially in forest tree populations (Brown et al. 2004; Neale and Savolainen 2004). Although linkage relationships between markers and QTLs can be established within pedigrees resulting from controlled crosses, historical recombination between markers and the QTL within populations means that QTL marker relationships must be reestablished in each new pedigree examined, and are completely uncertain in unrelated individuals taken from natural breeding populations.

The molecular analysis of plants often focused on the single gene level. But the recent technological advances have changed this paradigm. The way the genes and genetic information are organized within the genome, the methods of collecting and analyzing this information, and the determination of their biological functionality is referred to as genomics. Genomic approaches are permeating every aspect of plant biology, and since they rely on DNA-coded information, they expand molecular analyses from a single to a multispecies level. Plant genomics is reversing the previous paradigm of identifying genes behind biological functions and instead focuses on finding biological functions behind genes. It also reduces the gap between phenotype and genotype. This introductory chapter overviews two main sections: first, the current understanding of genomes, their genetic structure at the inter- and intraspecies level, and how whole genomes are sequenced; and second, on finding the biological and functional significance of DNA sequence. It is also worthwhile to note that these technologies, though extensively used in agricultural species, are only used in forest tree species research. Except for some tropical (avocado, mango, and papaya) and temperate (apple, *Prunus*, and *Pyurus*) fruit species, these techniques are not extensively used in other tree crops.

## Genetic Markers and Population Genetics

Molecular genetic markers have been extremely useful for tree population genetics, a discipline supporting basic research on the evolution of species and populations, and supporting applications ranging from tree improvement to conservation and restoration. Molecular markers have been used to estimate population parameters including population structure, gene flow, hybridization, migration, mating systems, and inbreeding. Knowledge of these attributes can be used to guide applications for management and conservation. For example, existing marker technologies can be used to determine levels of genetic diversity and inbreeding, two factors indicative of adaptive potential, which can help identify populations at risk. Existing markers can determine taxonomic relationships, a crucial component of establishing the legal basis for protection of endangered plant species. Contamination by nonlocal seed sources can erode the local adaptation of a population, and can potentially be detected using existing marker technology.

A major limitation of currently available markers is that they are neutral, meaning they are not within the actual genes that play a causative role in determining traits of interest. In addition, recombination and low linkage disequilibrium in tree populations means that linkage relationships between markers and alleles of genes controlling phenotypic traits are not consistent among individuals. This is a limiting factor for the application of marker technology to conservation and restoration applications because the markers have little or no predictive value for evaluating adaptive genetic attributes.

Significant sequence resources in the form of expressed sequence tags are available for numerous trees (http://plantta.tigr.org), with the largest conifer resource being >78,000 transcript assemblies for *Pinus taeda* (loblolly pine). Notably, these resources are being expanded through resequencing of alleles in support of association genetic studies (see below). Examples of additional resources for forest genomics include microarray resources (Abbott et al. 2008), proteomics (Lara et al. 2009), gene tagging and mutant collections (Layne and Bassi 2008), and ecotilling (Barkley and Wang 2008). For several angiosperms, transformation systems have been established that enable assessment of gene function using various strategies, including knockdown using RNAi (Enrique et al. 2011) or synthetic miRNAs (Song et al. 2010), or introduction of mutations into the amino acid sequence (Pillitteri et al. 2004).

Currently, association genetic studies in trees require a survey sequencing of alleles of candidate genes within a population to identify single nucleotide polymorphisms (SNPs) that define unique gene alleles. SNP genotypes and phenotypes are then measured for individuals sampled from the population, enabling testing for statistical association between SNP genotypes and phenotypes. Furthermore, linkage disequilibrium decays rapidly within a few hundred base pairs in both pine (Gonzalez-Martinez et al. 2006) and aspen (Ingvarsson 2005). As a result, an SNP with significant association with a phenotypic trait is likely to be close to or in the gene influencing the phenotype. This allows knowledge of gene function to be considered in understanding the genetic mechanisms regulating the trait being evaluated.

## Genetic Structure of Plant Genomes

Plant genomes are best described in terms of genome size, gene content, extent of repetitive sequences, and polyploidy/duplication events. Although plants also possess mitochondrial and chloroplast genomes, their nuclear genome is the largest and most complex. There is extensive variation in nuclear genome size (Table 1.1) without obvious functional significance of such variation (Rafalski 2002).

Plant genomes contain various repetitive sequences and retrovirus-like retrotransposons containing long terminal repeats and other retroelements, such as long interspersed nuclear elements and short-interspersed nuclear elements (Kumar and Bennetzen 1999). Retroelement insertions contribute to the large difference in size between collinear genome segments in different plant species and to the 50% or

**Table 1.1** Nuclear genome size in plants

| Common name | Nuclear genome size[a] |
|---|---|
| Wheat | 15,966 |
| Onion | 15,290 |
| Garden pea | 3,947 |
| Corn | 2,292 |
| Asparagus | 1,308 |
| Tomato | 907 |
| Sugar beet | 758 |
| Apple | 743 |
| Common bean | 637 |
| Cantaloupe | 454 |
| Grape | 483 |
| Man | 2,910 |

[a]Expressed in megabases (1 Mb = 1,000,000)

more difference in total genome size among species with relatively large genomes, such as corn. They contribute a smaller percentage of genome size in plants with smaller genomes such as *Arabidopsis* (The *Arabidopsis* Genome Initiative 2000). If other repetitive sequences are accounted for, then corn genome is comprised of over 70% repetitive sequences and 5% protein-encoding regions (Meyers et al. 2001).

It is widely accepted that 70–80% of flowering plants are the product of at least one polyploidization event (Barnes 2002). Many economically important plant species, such as corn, wheat, potato, and oat, are either ancient or more recent polyploids, comprising more than one, and in wheat three different, homologous genomes within a single species. Duplicated segments also account for a significant fraction of the rice genome. About 60% of the *Arabidopsis* genome is present in 24 duplicated segments, each more than 100 kilobases (kb) in size (Bevan et al. 2001). Ancestral polyploidy contributes to create genetic variation through gene duplication and gene silencing. Genome duplication and subsequent divergence is an important generator of protein diversity in plants.

## *Model Plant Species*

Model organisms (*Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae*) provide genetic and molecular insights into the biology of more complex species. Since the genomes of most plant species are either too large or too complex to be fully analyzed, the plant scientific community has adopted model organisms. They share features such as being diploid and appropriate for genetic analysis, being amenable to genetic transformation, having a (relatively) small genome and a short growth cycle, having commonly available tools and resources, and being the focus of research by a large scientific community. Although the advent of tissue culture techniques fostered the use of tobacco and petunia, the species now

used as model organisms for mono- and dicotyledonous plants are rice (*Oryza sativa*) and *Arabidopsis* (*Arabidopsis thaliana*), respectively.

*Arabidopsis*, a small Cruciferae plant without agricultural use, sets seed in only 6 weeks from planting, has a small genome of 120 Megabases (Mb), and only five chromosomes. There are extensive tools available for its genomic analysis, whole genome sequence, Expressed Sequence Tags (ESTs) collections, characterized mutants, and large populations mutagenized with insertion elements (transposons or the T-DNA of *Agrobacterium*). *Arabidopsis* can be genetically transformed on a large scale with *Agrobacterium tumefaciens* and biolistics. Other tools available for this model plant are saturated genetic and physical maps.

Unlike *Arabidopsis*, rice is one of the world's most important cereals. More than 500 million tons of rice is produced each year, and it is the staple food for more than half of the world's population. There are two main rice subspecies. *Japonica* is mostly grown in Japan, while *indica* is grown in China and other Asia-Pacific regions. Rice also has very saturated genetic maps, physical maps, whole genome sequences, as well as EST collections pooled from different tissues and developmental stages. It has 12 chromosomes, a genome size of 420 Mb, and like *Arabidopsis*, it can be transformed through biolistics and *A. tumefaciens*. Efficient transposon-tagging systems for gene knockouts and gene detection have not yet become available for saturation mutagenesis in rice, although some recent successes have been reported.

## *Maps*

### Genetic Maps

The development of molecular markers has allowed for constructing complete genetic maps for most economically important plant species. They detect genetic variation directly at the DNA level. A myriad of molecular marker systems are available, yet their description lies beyond the scope of this paper. A genetic map represents the ordering of molecular markers along chromosomes as well as the genetic distances, generally expressed as centiMorgans (cM), existing between adjacent molecular markers. Genetic maps in plants have been created from many experimental populations, but the most frequently used are $F_2$, backcrosses, and recombinant inbred lines. Although longer to develop, recombinant inbred lines offer a higher genetic resolution and practical advantages. Once a mapping population has been created, it takes only few months to produce a genetic map with a 10-cM resolution. Genetic maps contribute to the understanding of how plant genomes are organized, and once available, they facilitate the development of practical applications in plant breeding, such as the identification of Quantitative Trait Loci and Marker-Assisted Selection. Most economically important plant traits such as yield, plant height, and quality components exhibit a continuous distribution rather than discrete classes and are regarded as quantitative traits. These traits are

controlled by several loci each of small effect, and different combinations of alleles at these loci can give different phenotypes.

Quantitative Trait Loci analysis refers to the identification of genomic regions associated with the phenotypic expression of a given trait. Once the location of such genomic regions is known, they can be assembled into designer genotypes, that is, individuals carrying chromosomic fragments associated with the expression of a given phenotype. The most important feature of Marker-Assisted Selection is that once a molecular marker genetically linked to the expression of a phenotypically interesting allele has been detected, an indirect selection for such allele based upon the detection of the molecular marker can be accomplished, since little or any genetic recombination will occur between them. Therefore, the presence of the molecular marker will always be associated with the presence of the allele of interest.

Genetic maps are also an important resource for plant gene isolation, as once the genetic position of any mutation is established, it is possible to attempt its isolation through positional cloning (Campos de Quiroz et al. 2000). Furthermore, genetic maps help establish the extent of genome collinearity and duplication between different species.

## Physical Maps

Although genetic maps provide much-needed landmarks along chromosomes, they are still too far apart to provide an entry point into genes, since even in model plants the kilobases per centiMorgan (kb/cM) ratio is large, from 120 to 250 kb/cM in *Arabidopsis* and between 500 and 1.500 kb/cM in corn. Therefore, a 1-cM interval may harbor ~30 to 100 or even more genes. Physical maps bridge such gaps, representing the entire DNA fragment spanning the genetic location of adjacent molecular markers.

Physical maps can be defined as a set of large insert clones with minimum overlap encompassing a given chromosome. First-generation physical maps in plants were based on YACs (Yeast Artificial Chromosomes). Chimerism and stability issues, however, dictated the development of low copy, *E. coli*-maintained vectors such as Bacterial Artificial Chromosomes (BACs) and P1-derived artificial chromosomes. Although BAC vectors are relatively small (molecular weight of BAC vector pBeloBAC11 is 7.4 kb for instance), they carry inserts between 80 and 200 kb on average and possess traditional plasmid selection features such as an antibiotic resistance gene and a polycloning site within a reporter gene allowing insertional inactivation. BAC clones are easier to manipulate than yeast-based clones. Once a BAC library is prepared, clones are assembled into contigs using fluorescent DNA fingerprint technologies and matching probabilities. Physical and genetic maps can be aligned, bringing along continuity from phenotype to genotype. Furthermore, they provide the platform clone-by-clone sequencing approaches rely upon. Physical maps provide the bridge needed between the resolution achieved by genetic maps and that needed to isolate genes through positional cloning.

## *Genome Collinearity/Genome Evolution*

A remarkable feature of plant genomics is its ability to bring together more than one species for analysis. The comparative genome mapping of related plant species has shown that the organization of genes is highly conserved during the evolution of members of taxonomic families. This has led to the identification of genome collinearity between the well-sequenced model crops and their related species (e.g., *Arabidopsis* for dicots and rice for monocots). Collinearity overrides the differences in chromosome number and genome size and can be defined as conservation of gene order within a chromosomal segment between different species. A related concept is synteny, which refers to the presence of two or more loci on the same chromosome, regardless of whether they are genetically linked or not.

Collinear relationships have been observed among cereal species (corn, wheat, rice, barley), legumes (beans, peas, and soybeans), pines, and *Cruciferae* species (canola, broccoli, cabbage, *Arabidopsis thaliana*). Recently, the first studies at the gene level have demonstrated that microcollinearity of genes is less conserved; small-scale rearrangements and deletions complicate microcollinearity between closely related species. For instance, although a 78-kb genomic sequence of sorghum around the locus *adh1* and its homologous genomic fragment from maize showed considerable microcollinearity and the fact that they share nine genes in perfect order and transcriptional direction, five additional, unshared genes reside in this genomic region (Tikhonov et al. 1999).

Comparing sequences of soybean and *Arabidopsis* demonstrated partial homology between two soybean chromosomes and a 25-cM section of chromosome 2 from *Arabidopsis* (Lee et al. 2001). Although such relationships need to be assessed on a case-by-case basis, they reflect the value *Arabidopsis* and other model species offer to economically important species.

Collinearity has also been established between rice and most cereal species, allowing the use of rice for genetic analysis and gene discovery in genetically more complex species, such as wheat and barley (Shimamoto and Kyozuka 2002). A comparison of rice and barley DNA sequences from syntenic regions between barley chromosome 5H and rice chromosome 3 revealed the presence of four conserved regions, containing four predicted genes. General gene structure was largely conserved between rice and barley (Dubcovsky et al. 2001). A similar comparison between corn and rice, based on 340 kb around loci *adh1* and *adh2*, showed five collinear genes between the two species, as well as a possible translocation on *adh1*. Rice genes similar to known disease-resistant genes showed no cross-hybridization with corn genomic DNA, suggesting sequence divergence or their absence in maize (Tarchini et al. 2000). There are even reports of collinearity across the mono-dicotyledoneous division involving *Arabidopsis* and cereals, which diverged as far back as 200 million years ago (Mayer et al. 2001) Exploiting collinearity helps to establish cross-species genetic links and also aids in the extrapolation of information from species with simpler genomes (i.e., rice) to genetically complex species (corn, wheat). Furthermore, it reflects the power of genomics to integrate genetic information across species.

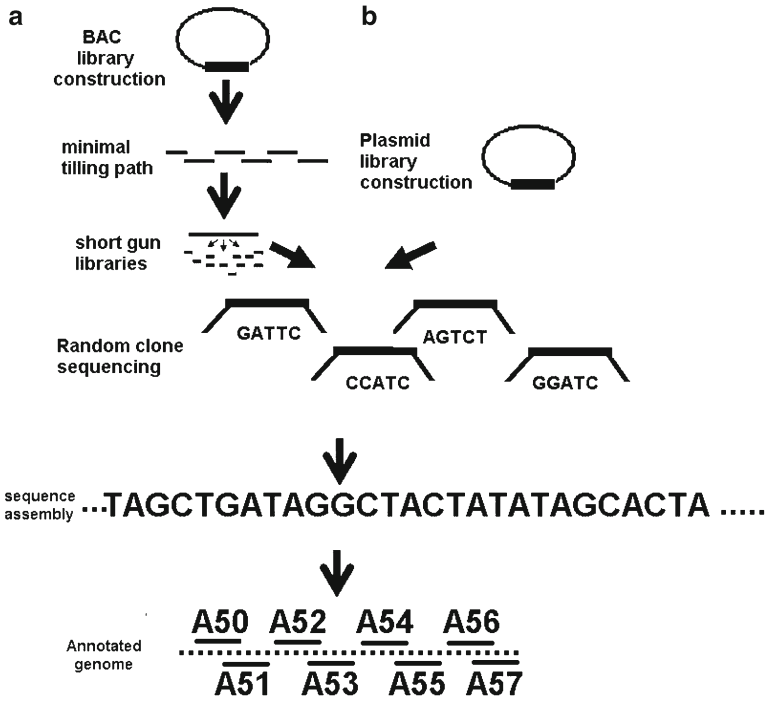


**Fig. 1.1** Approaches of large scale sequencing (**a**) clone-by-clone strategy and (**b**) short gun strategy

## *Whole Genome Sequencing*

Genetic and physical maps at the inter- or intraspecies level represent a key layer of genomic information. However, sequence data represents the ultimate level of genetic information. Three major breakthroughs have allowed the sequencing of complete genomes: (1) The development of fluorescence-based DNA sequencing methods that provide at least 500 bases per read; (2) The automation of several processes such as picking and arraying bacterial subclones, purification of DNA from individual subclones, and sample loading among others; and (3) The development of software and hardware able to handle massive amounts (gigabytes) of data points.

There are two main approaches to large-scale sequencing (Fig. 1.1). In clone-by-clone strategies (Fig. 1.1a), large insert libraries, such as those based on BAC clones, are used as sequencing templates, and inserts are arranged into contigs using diverse fingerprinting methods to establish minimal tiling paths. Sequence-tagged connectors extracted from large insert clones as well as FISH (fluorescence in situ hybridization) and optical mapping are used to extend contigs and close gaps (Marra et al. 1997). BAC clones from sequence-ready contigs are then fragmented into plasmid or M13 vector-based shotgun libraries with insert sizes of ~1–3 kb. Using more than one vector system reduces cloning bias issues. Sequencing efforts are tailored to the

degree of coverage required. For instance, for a fivefold coverage, and assuming 500 base pairs (bp) per sequencer reading, 800 clones are sequenced to cover an 80-kb BAC clone. Finished sequences are those obtained at a ~8–10-fold coverage and provide >99.99% accuracy, whereas working draft sequences are attained at a ~3–5-fold coverage. It is important to note, however, that even working draft sequences provide an enormous amount of information, and even shotgun approaches rely to some extent on clone-by-clone information.

After sequencing is concluded, the DNA data are used to reassemble BAC clones. Base calling programs assigning quality scores to each read base such as Phred (Ewing et al. 1998), sequence assembly programs such as Phrap (Gordon et al. 1998), and graphical viewing tools are used to achieve such assembly. The finishing of the sequence then ensues, which can be done in part manually or with finishing software, such as Autofinish (Gordon et al. 2001).

Annotation, or the process of identifying start and stop codons and the position of introns that permits the prediction of biological function from DNA sequence, proceeds through three main steps. The first is to use gene finders like Xgrail (Uberbacher and Mural 1991) or others based on generalized hidden Markov models, such as GeneMark.hmm (Lukashin and Borodovsky 1998) and GenScan (Burge and Karlin 1997), specifically developed to recognize *Arabidopsis* genes. In the second step, sequences are aligned to protein and EST databases; and finally, putative functions are assigned to each gene sequence. Successful annotation processes often combine different software and manual inspection.

In shotgun approaches (Fig. 1.1b), which have been successfully used to sequence many microorganisms and *D. melanogaster*, small insert libraries are prepared, and randomly selected inserts are sequenced until a ~5-fold or higher coverage is reached. Sequences are then assembled, gaps are identified and closed, and finally, annotation is conducted. Shotgun sequencing does not rely upon the availability of minimal tiling paths and, therefore, reduces the cost and effort required to obtain whole genome sequences. Nevertheless, they require an enormous amount of computational power to assemble a large number of random sequences into a small number of contigs. Furthermore, the ultimate quality of large genomes that have been shotgun-sequenced may not be as high as that achievable using the clone-by-clone approach. Because of a high content of long and highly conserved repetitive sequences, including retrotransposons, shotgun sequencing of plant genomes may pose special challenges.

## *Reverse Genetics*

Traditional genetic analysis aims to identify the DNA sequences associated with a given phenotype. Reverse genetics determines the function of a gene for which the sequence is known, by generating and analyzing the phenotype of the corresponding knockout mutant (Maes et al. 1999). Unlike yeast, in which gene disruption is available through homologous recombination, transposon and T-DNA tagging are the

best methods available for developing mutagenized plant populations suitable for reverse genetics studies (Pereira 2000). There are several mutagenized populations in *Arabidopsis* suited for reverse genetics studies. A European consortium is developing heterologous systems for rice based on the Ac element from corn (Greco et al. 2001). There are also proprietary populations, such as Pioneer Hi-Bred International's Trait Utility System for Corn (TUSC), mutagenized with the high copy Mu element (Multani et al. 1998). Using high copy elements makes it possible to use smaller populations to ensure that tagged mutants will be found for most genes.

There are two main possibilities for identifying tagged genes at insertion sites. For unknown genes, sequences flanking the insertion can be obtained through inverse polymerase chain reaction (PCR) (Ochman et al. 1988) or thermal assimetric interlaced PCR (Liu and Whittier 1995), whereas for insertions in genes of known sequence, it is possible to amplify and clone the sequence of interest through PCR using gene-specific and insertion-specific primers. Since in the latter case it is common to analyze thousands of plants, PCR-based screening is arranged into three-dimensional pools that allow the unequivocal identification of tagged individuals. Large databases of characterized insertion sites are becoming available that will further ease the use of insertion elements to isolate useful genes (Tissier et al. 1999).

Although several genes have been isolated through reverse genetic approaches, two main factors have limited their wider application. First, many genes are functionally redundant, as even species with simple genomes such as *Arabidopsis* carry extensive duplications, and second, mutations in many genes may be highly pleiotropic, which can mask the role of a gene in a specific pathway (Springer 2000). Nevertheless, reverse genetics is considered to be a major component of the functional genomics toolbox, and it plays an important role in assigning biological functions to genes discovered through large-scale sequencing programs. Transposon tagging provides an excellent alternative to isolate tagged genes that exhibit relatively simple inheritance.

Gene traps refer to another application of transposons that responds to regulatory sequences at the site of insertion. Depending on the sequences engineered, they can be classified as reporter traps, enhancer traps, or gene traps. Since they rely on reporter gene expression, mutant phenotypes are not required, and they have been valuable in isolating tissue and cell-specific sequences (Springer 2000).

## *Transcriptional Profiling*

While molecular biology generally analyzes one or a few genes simultaneously, recent developments allow the parallel analysis of thousands of genes. This area of genomics involves the study of gene expression patterns across a wide array of cellular responses, phenotypes, and conditions. The expression profile of a developmental stage or induced condition can identify genes and coordinately regulated pathways and their functions. This produces a more thorough understanding of the underlying biology (Quackenbush 2001).

There are several systems available to analyze the parallel expression of many genes, such as macroarrays (Desprez et al. 1998), microarrays (Schena et al. 1995), and serial analysis of gene expression (SAGE) (Velculescu et al. 1995), which consists of identifying short sequence tags from individual transcripts, their concatenation, sequencing, and subsequent digital quantitation. SAGE provides expression levels for many transcripts across different stages of development.

There are open and closed transcriptional profiling systems. Open technologies survey a large number of transcripts and analyze their levels between different samples, but the identity of the genes involved is not known *a priori*. One example of such a system is the GeneCalling technology (Bruce et al. 2000). Another open system is provided by massively parallel sequence signatures (MPSS), where microbeads are used to construct libraries of DNA templates and create hundreds of thousands of gene signatures (Brenner et al. 2000).

Closed systems, on the other hand, analyze genes that have been previously characterized. They include most of the diverse microarray systems available, and these are based on the specific hybridization of labeled samples to spatially separate immobilized nucleic acids, thus enabling the parallel quantification of many specific mRNAs. It is important to select the system at the onset of any transcriptional profiling study and stay with it.

The focus here is on microarrays. In microarray experiments, DNA samples corresponding to thousands of genes of interest are immobilized on a solid surface such as glass slides in a regular array. The immobilized sequences are usually referred to as probes. RNA samples (or their cDNA derivatives) from biological samples under study are hybridized to the array and are referred to as the target. Labeling with fluorescent dyes with different excitation and emission characteristics allows the simultaneous hybridization of two contrasting targets on a single array (Aharoni and Vorst 2001).

Microarray applications are broadly classified as expression-specific and genome-wide expression studies. In expression-specific studies, they are used as a functional genomics tool to address the biological significance of genes discovered through large-scale sequencing, as well as a means to understanding the genetic networks explaining biological processes or biochemical pathways. The value of using microarrays to identify novel response genes has been demonstrated by studying the gene expression patterns during corn embryo development (Lee et al. 2002), the response to drought and cold stresses (Seki et al. 2001), herbivory (Arimura et al. 2000), and nitrate treatments (Wang et al. 2000).

When addressing a specific pathway or biological process, it is useful to include genes beyond those of apparent interest, since over-specific microarrays would not be able to address genetic interactions with other biological processes. This principle revealed previously unexpected relationships between low soil phosphate levels and cold acclimation in *Arabidopsis* (Hurry et al. 2000). Genes obtained from the transcriptional analysis of plant responses to stress are of particular relevance for transgenic approaches, as thoroughly reviewed by Dunwell et al. (2001).

Genome-wide arrays are mostly designed for model organisms such as *Arabidopsis* or rice, as there are many genes available to select from, either as clones or as annotated genomic sequences for model species. They are also available to

species such as corn that have extensive EST collections. This enabling technology is an immediate and direct result of large-scale sequencing projects. It is expected that microarrays covering most of the *Arabidopsis* genes will become available in 2003. Genome-wide expression profiles are the ultimate tool to integrate all genes existing in an organism into a series of experiments. They also help to elucidate the coordinate expression of different genetic networks and document how changes in one would impact others. It is expected that such genome-wide approaches will be particularly useful in identifying new regulatory sequences and master switches that affect distinct but apparently unrelated genetic networks.

Transcriptional profiling technologies play a central role in predicting gene function since sequence comparison alone is insufficient to infer function. They also help to detect phenomena such as gene displacement – nonhomologous genes coding for proteins that serve the same function – and gene recruitment – genes with identical sequences coding for completely different functions (Noordewier and Warren 2001).

Unlike animals, plants cannot move and have developed exquisite mechanisms to cope with changing environmental conditions and biotic challenges, since these directly or indirectly affect most biological processes occurring in plants. Therefore, a significant proportion of the information gathered by specific and genome-wide transcription profiling processes should have practical applications and facilitate the development of plants more resilient to biotic and abiotic stimuli.

## Concluding Remarks

The current understanding of plant biology is limited to whole organism biology and its gene functions. Genomics adds another level of understanding to plant biology through the integrated analysis of different species. The large number of genes handled simultaneously by genomics sets a new paradigm in plant biology, since it allows the genetic integration of diverse processes, tissues, and organisms.

Finally, genomics is the ultimate interdisciplinary approach, as it covers the entire spectrum from DNA sequencing to field-based research. The integrated endeavor of genetics, biology, bioinformatics, molecular biology, engineering, microbiology, and related fields will extensively benefit mankind.

Genomics of trees, however, is in its infancy. Concerted efforts by tree biotechnologists will only benefit this paradigm. Except breakthroughs in some temperate and tropical fruit trees, they are to be fully exploited though studies on genomics.

## References

Abbott DA, Suir E, van Maris AJ, Pronk JT (2008) Physiological and transcriptional responses to high concentrations of lactic acid in anaerobic chemostat cultures of *Saccharomyces cerevisiae*. Appl Environ Microbiol 18:5759–5768

Adams J et al (2000) The case for genetic engineering of native and landscape trees against introduced pests and diseases. Conserv Biol 16:874–879

Aharoni A, Vorst O (2001) DNA microarrays for functional plant genomics. Plant Mol Biol 48:99–118 [Links]

Arimura G, Tashiro K, Kuhara S, Nishikova T, Ozawa R, Takabayashi J (2000) Gene responses in bean leaves induced by herbivory and by herbivore-induced volatiles. Biochem Biophys Res Commun 277:305–310 [Links]

Barkley NA, Wang ML (2008) Application of TILLING and EcoTILLING as reverse genetic approaches to elucidate the function of genes in plants and animals. Curr Genomics 9(4):212–226

Barnes S (2002) Comparing *Arabidopsis* to other flowering plants. Curr Opin Plant Biol 5:128–133

Bevan M, Mayer M, Mayer K, White O, Eisin JA, Preuss D, Bureau T, Salzberg S, Mewes HW (2001) Sequence and analysis of the *Arabidopsis* genome. Curr Opin Plant Biol 4:105–110

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo SJ, Mccurdy S, Foy M, Ewan M, Roth R, George S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, Dubridge RB, Kirchner J, Fearon K, Mao J, Corcoran NK (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634

Brown GR et al (2004) Nucleotide variation and linkage disequilibrium in loblolly pine. Proc Natl Acad Sci USA 101:15255–15260

Bruce W, Folkerts O, Garnaat C, Crasta O, Roth B, Bowen B (2000) Expression profiling of the maize flavonoid pathway genes controlled by estradiol-inducible transcription factors CRC and P. Plant Cell 12:65–79

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268:78–94

Campos de Quiros H, Magrath R, McCallum D, Kroymann J, Schnabelrauch D, Mitchell-Olds T, Mithen R (2000) α-Keto acid elongation and glucosinolate biosynthesis in *Arabidopsis thaliana*. Theor Appl Genet 101:429–437

Desprez T, Amselem J, Caboche M, Hofte H (1998) Differential gene expression in *Arabidopsis* monitored using cDNA arrays. Plant J 14:643–652

DiFazio SP, Slavov GT, Burczyk J, Leonardi S, Strauss SH (2004) Gene flow from tree plantations and implications for transgenic risk assessment. In: Forest Biotechnology for the 21st Century (eds. Walter C & Carson M), Research Signpost, Kerala, India. pp. 405–422

Dubcovsky J, Ramakrishna W, Sanmiguel P, Busso C, Yan L, Shiloff B, Bennetzen JL (2001) Comparative sequence analysis of colinear barley and rice Bacterial Artificial Chromosomes. Plant Physiol 125:1342–1353

Dunwell JM, Moya-Leon MA, Herrera R (2001) Transcriptome analysis and crop improvement: a review. Biol Res 34(3–4):153–164

Enrique R, Siciliano F, Favaro MA, Gerhardt N, Roeschlin R, Rigano L, Sendin L, Castagnaro A, Vojnov A, Marano MR (2011) Novel demonstration of RNAi in citrus reveals importance of citrus callose synthase in defence against *Xanthomonas citri* subsp. *Citri*. Plant Biotechnol J 9:394–407

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175–185

Gonzalez-Martinez SC et al (2006) DNA sequence variation and selection of tag SNPs at candidate genes for drought-stress response in *Pinus taeda* L. Genetics 172:1915–1926

Gordon D, Abaijan C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Gordon D, Desmarais C, Green P (2001) Automated finishing with Autofinish. Genome Res 11:614–625

Greco R, Ouwerkerk PBF, Sallaud C, Kohli A, Colombo L, Puigdomenech P, Guiderdoni E, Christou P, Hoge JHC, Pereira A (2001) Transposon insertional mutagenesis in rice. Plant Physiol 125:1175–1177

Herrera S (2005) Struggling to see the forest though the trees. Nat Biotechnol 2:165–167

Hurry V, Strand A, Furbank R, Stitt M (2000) The role of inorganic phosphate in the development of freezing tolerance and the acclimatization of photosynthesis to low temperature is revealed by the pho mutants of *Arabidopsis thaliana*. Plant J 24:383–396

Ingvarsson PK (2005) Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). Genetics 169:945–953

Jermstad KD et al (2003) Mapping of quantitative trait loci controlling adaptive traits in Douglas fir. III. Quantitative trait lociby-environment interactions. Genetics 165:1489–1506

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Ann Rev Genet 33:479–532

Lara MV, Borsani J, Budde CO, Lauxmann MA, Lombardo VA, Murray R, Andreo CS, Drincovich MF (2009) Biochemical and proteomic analysis of 'Dixiland' peach fruit (*Prunus persica*) upon heat treatment. J Exp Bot 60:4315–4333

Layne DR, Bassi D (2008) The peach: botany, production and uses (CABI). CABI 30 Nov 2008, 848 pp. ISBN: 1845933869 PDF 27.1 MB

Lee JM, Grant D, Vallejos CE, Shoemaker RC (2001) Genome organization in dicots II *Arabidopsis* as a ëbridging speciesí to resolve genome evolution events among legumes. Theor Appl Genet 103:765–773

Lee JM, Williams ME, Tingey SV, Rafalski JA (2002) DNA array profiling of gene expression changes during maize embryo development. Funct Integr Genomics 2:13–27

Liu Y, Whittier R (1995) Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25:674–681

Lukashin AV, Borodovsky M (1998) GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 26:1107–1115

Maes T, De Keukeleire P, Gerats T (1999) Plant tagnology. Trends Plant Sci 4:90–96

Marra M, Kucaba T, Dietrich N, Green E, Brownstein WRK, Mcdonald K, Hillier L, Mcpherson J, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res 7:1072–1084

Mayer K, Murphy G, Tarchini R, Wambutt R, Volckaert G, Pohl T, Düsterhöft A, Stiekema W, Entian K-D, Terryn N, Lemcke K, Haase D, Hall C, van Dodeweerd A-M, Tingey S, Mewes H-W, Bevan MW, Bancroft I (2001) Conservation of microstructure between a sequenced region of the genome of rice, multiple segments of the genome of *Arabidopsis thaliana*. Genome Res 11:1167–1174

Meyers B, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Multani D, Meeley RB, Paterson AH, Gray J, Briggs SP, Johal GS (1998) Plant-pathogen microevolution: molecular basis for the origin of a fungal disease in maize. Proc Natl Acad Sci USA 95:1686–1691

Neale DB, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9:325–330

Noordewier MO, Warren PV (2001) Gene expression microarrays, the integration of biological knowledge. Trends Biotechnol 19:412–415

Ochman H, Gerber A, Hart D (1988) A genetic application of an inverse polymerase chain reaction. Genetics 120:621–623

Pereira A (2000) A transgenic perspective on plant functional genomics. Transgenic Res 9:245–260

Pillitteri LJ, Lovatt CJ, Walling LL (2004) Isolation and characterization of a *TERMINAL FLOWER* homolog and its correlation with juvenility in citrus. Plant Physiol 135:1540–1551

Quackenbush J (2001) Computational analysis of microarray data. Nat Rev Genet 2:418–427

Rafalski AJ (2002) Plant genomics: present state and a perspective on future developments. Brief Fundam Genomics Proteomics 1:1–15

Schena M, Shalon D, Davis RW, Brown P (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Seki M, Narusaka M, Abe H, Kasuga M, Yamaguchi-Shinozaki K, Carninci P, Hayashizaki Y, Shinozaki K (2001) Monitoring the expression pattern of 1300 *Arabidopsis* genes under drought, cold stresses by using a full-length cDNA microarray. Plant Cell 13:61–72

Shimamoto K, Kyozuka J (2002) Rice as a model for comparative genomics of plants. Annu Rev Plant Biol 53:399–419

Song C, Cao X, Nicholas KK, Wang C, Li X, Wang X, Fang J (2010) Extraction of low molecular weight RNA from *Citrus trifolita* tissues for microRNA Northern blotting and reverse transcriptase polymerase chain reaction (RTPCR). Afr J Biotechnol 9:8726–8730

Springer PS (2000) Gene traps: tools for plant development and genomics. Plant Cell 12:1007–1020

Tarchini R, Biddle P, Winel R, Tingey S, Rafalski A (2000) The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 region reveals interrupted colinearity with maize chromosome 4. Plant Cell 12:381–391

The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis* thaliana. Nature 408:796–815

Tikhonov A, Sanmiguel P, Nakajimanina Y, Gorenstein M, Bennetzen JL, Avramova Z (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. Proc Natl Acad Sci USA 96:7409–7414

Tissier AF, Marillonnet S, Klimyuk V, Patel K, Torres MA, Murphy G, Jones JD (1999) Multiple independent defective suppressor-mutator transposon insertions in *Arabidopsis*: a tool for functional genomics. Plant Cell 11:1841–1852

Uberbacher EC, Mural RJ (1991) Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach. Proc Natl Acad Sci USA 88:11261–11265

van Frankenhuyzen K, Beardmore T (2004) Current status and environmental impact of transgenic forest trees. Canadian Journal of Forest Res 34(6):1163–1180

Velculescu V, Zhang L, Vogelstein B, Kinzler K (1995) Serial analysis of gene expression. Science 270:484–487

Wang R, Guegler K, Labrie Samuel T, Crawford NM (2000) Genomic analysis of a nutrient response in *Arabidopsis* reveals diverse expression patterns and novel metabolic and potential regulatory genes induced by nitrate. Plant Cell 12:1491–1509

# Chapter 2
# Bioinformatics Techniques for Understanding and Analyzing Tree Gene Expression Data

**Lewis Lukens and Gregory Downs**

**Abstract** There is great interest in enhancing our understanding of the molecular bases of tree biological processes using genomic techniques. One such technique is transcriptional profiling that assays the transcript abundance of thousands of genes. The analyses of these inventories of gene expression help explain the genetic diversity of trees and trees' responses to different developmental stages and environmental conditions. In this chapter, we describe key approaches for collecting transcriptome data and the tree genomic resources available for this data's use and interpretation. We define the factors that cause gene transcript abundances to vary and elucidate how to quantify these factors' effects. We also describe approaches to identify co-regulated genes and to assign functions to genes and groups of genes. Finally, we suggest future directions for tree transcriptome analyses.

**Keywords** Bioinformatics • Expression analysis • Microarray • High-throughput sequencing • Transcriptome • Gene expression variation • Gene co-regulation • Functional annotation

## Introduction

Trees are a critical component of our environment covering a substantial proportion of the earth's land area. They are taxonomically diverse. Although characteristic of the gymnosperms, tree species are found in over 100 plant families, indicating frequent evolution of the tree growth habit. Finally, trees have major economic importance. One key attribute of trees is woodiness. Wood is a major renewable natural resource for the timber, fiber, and bioenergy industries. Tree species are also the

L. Lukens, Ph.D. (✉) • G. Downs, B.Sc.
Department of Plant Agriculture, University of Guelph,
50 Stone Road East, Guelph, ON, Canada N1G2W1
e-mail: llukens@uoguelph.ca

R.J. Schnell and P.M. Priyadarshan (eds.), *Genomics of Tree Crops*,
DOI 10.1007/978-1-4614-0920-5_2, 