Jonathan F. Wendel
Editor-in-chief

Johann Greilhuber
Jaroslav Doležel
Ilia J. Leitch
Editors

Plant Genome Diversity

Volume 1: Plant Genomes, Their Residents, and Their Evolutionary Dynamics

Springer

Plant Genome Diversity Volume 1

Jonathan F. Wendel
Editor-in-chief

Johann Greilhuber · Jaroslav Doležel ·
Ilia J. Leitch
Editors

Plant Genome Diversity Volume 1

Plant Genomes, Their Residents, and Their Evolutionary Dynamics

Springer

Editor-in-chief
Jonathan F. Wendel
Department of Botany
Iowa State University
Ames, Iowa
USA

Editors
Johann Greilhuber
Department of Systematic and Evolutionary Botany
Faculty Center Botany
University of Vienna
Vienna
Austria

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

Ilia J. Leitch
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, Surrey
United Kingdom

ISBN 978-3-7091-1129-1 ISBN 978-3-7091-1130-7 (eBook)
DOI 10.1007/978-3-7091-1130-7
Springer Wien Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012935228

Printed on acid-free paper

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

Preface

Ever since Darwin, biologists have been interested in understanding the intricacies of natural variation patterns and the evolutionary forces that shape this diversity. Although these dual objectives have long been central goals in evolutionary biology, they recently have assumed a new prominence arising from the development of breathtaking new genomic technologies and their application to natural systems. These technological leaps have invigorated the already thriving discipline variously referred to as *molecular evolution*, or *genome evolution*, or sometimes *evolutionary genomics*, empowering it with a vastly expanded insight into the diversity of genomes and their evolutionary dynamics. As a result, excitement in the discipline has never been higher. Notwithstanding this vibrancy and the attendant meteoric rise in the number of published papers and professional journals devoted to molecular evolution, remarkably few books on the topic are available. To be sure, there are a number of classic texts, written more than a decade ago, just as the genomics revolution was ramping up, and a larger suite of hybrid books variously combining aspects of bioinformatics, genome evolution, population genetics, and methods of phylogenetic inference. To date, however, there exists no single modern treatment of what we have learned about the diversity and evolution of plant genomes and their various genomic residents.

It was to fill this void that the present project was initiated. Inspired by the seemingly ever-expanding pace of insights into the evolution of plant genomes, and motivated by the desire to provide for students and researchers a single point of entry into a burgeoning literature, we invited leading authorities in plant molecular evolution to participate in a project aimed at providing a comprehensive (but not encyclopedic) yet accessible introduction to the current state of the art in the field. This is accomplished here in a total of 16 chapters that collectively cover the discipline. Although these are arranged in a logical progression and are interconnected, each chapter also serves as a stand-alone introduction and review, thus providing a text that may flexibly be used by advanced undergraduate students, graduate students, and professionals in many fields in the plant sciences and beyond.

The volume appropriately begins (Flagel and Blackman, Chap. 1) with a review of the immense insights that have been gleaned from plant genome sequencing projects, as well as a prospective view of both the promises and challenges that lie ahead. This is followed by two complementary chapters on the primary constituents of plant genomes, namely, transposable elements (TEs); the first of these (Kejnovsky et al., Chap. 2) focuses on the diversity of TEs, their genomic ecology, and their role in genome size evolution, whereas the second (Slotkin et al., Chap. 3) reviews the remarkable role TEs play in genetic and epigenetic regulation, and as evolutionary fodder for the origin of novel genes and for chromosomal evolution. Perhaps the most obvious features of chromosomes are centromeres and telomeres, for which our knowledge regarding structure and evolution have been dramatically increased by genomic technologies, as reviewed by Hirsch and Jiang (centromeres, Chap. 4) and Siomos and Riha (telomeres, Chap. 5).

Having described the major structural features and organization of plant genomes, we turn our attention to smaller genomic residents, including small RNAs, for which Lee et al. (Chap. 6) present a synopsis of the diversity, regulatory roles, and evolution of the different classes of small RNAs. This is followed by a chapter on genic evolution, with a special focus on rate variation within and among lineages and the utility of this information for timing divergence events (Burleigh, Chap. 7), and on the detection and significance of conserved non-coding DNA (Subramaniam and Freeling, Chap. 8). Mowers et al. (Chap. 9) and Wolf (Chap. 10) offer timely reviews of the structure and evolutionary dynamics of plant mitochondrial and plastid genomes, respectively.

One of the key emergent realizations of the genomics era has been that plant genomes are replete with evidence of historical and ongoing duplications, large and small. Barker et al. (Chap. 11) review the processes that generate duplications as well as their longer term evolutionary outcomes, whereas Nieto-Feliner and Rossello (Chap. 12) present an update on a curious non-Mendelian consequence of sequence multiplicity, namely, sequence homogenization via one or more means of "concerted evolution". Paterson et al. (Chap. 13) further describe the consequences of genome duplication and divergence on longer-term colinearity and synteny relationships among divergent lineages. A final consequence of genome divergence, variation in base composition, is considered by Šmarda and Bureš (Chap. 14), who provide an overview of the phenomenon and its possible causative forces. We close with two chapters devoted to the vibrant new frontier of plant epigenomics, one (Zhang, Chap. 15) describing the epigenetic landscape in plants and the various forms of chromatin modification, and the second (Richards et al., Chap. 16) devoted to the evolutionary signification of epigenetic variation.

We are living in a tremendously exciting time to be a biologist, perhaps one that in the future will be thought of as having been a "golden era", replete with technological and conceptual breakthroughs. We hope that you find this volume evocative in this sense, as stimulating to read as it was to produce, and inspiring in the promise of its content.

Of course there are many people to thank for bringing this project to fruition. First and foremost are the many authors, who are experts in their field and hence are very busy people. Yet they willingly and generously set aside the time to imagine and create their contributions. To them I offer my sincere appreciation. I also offer thanks to many of my professional colleagues, who reviewed drafts of the manuscripts and offered numerous insights and suggestions for improvement. It should go without saying that this book would not have been possible without the vision of the publisher, Springer-Verlag, who recognized that this volume would fill an important and presently largely vacant niche, and for moving the project along expeditiously toward completion. Finally, I need to express my delighted gratitude to my co-editors of this soon-to-be two volume set, Ilia Leitch, Jaroslav Doležel, and especially Editor-in-Chief Johann Greilhuber, for their many modes of assistance throughout project conception and execution.

Ames, Iowa (United States) Jonathan Wendel

Contents

Contributions

A. J. Alverson Department of Biology, Indiana University, Bloomington, IN, USA, andy.alverson@gmail.com

M.S. Barker Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, USA, msbarker@email.arizona.edu

G. J. Baute Department of Botany, University of British Columbia, Vancouver, Canada, gregbaute@gmail.com

Benjamin K. Blackman Department of Biology, Duke University, Durham, NC, USA, bkb7@duke.edu

O. Bossdorf Institute of Plant Sciences, University of Bern, Bern, Switzerland

Petr Bureš Faculty of Science, Department of Botany and Zoology, Masaryk University, Brno, Czech Republic, bures@sci.muni.cz

J. G. Burleigh Department of Biology, University of Florida, Gainesville, FL, USA, gburleigh@ufl.edu

Cédric Feschotte Department of Biology, University of Texas, Arlington, TX, USA, cedric@uta.edu

Lex E. Flagel Department of Biology, Duke University, Durham, NC, USA, lex.flagel@duke.edu

M. Freeling Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA, freeling@berkeley.edu

Jennifer S. Hawkins Department of Biology, West Virginia University, Morgantown, WV, USA, jhawkins@uga.edu

C.D. Hirsch Department of Horticulture, University of Wisconsin-Madison, Madison, WI, USA, cdhirsch@wisc.edu

J. Jiang Department of Horticulture, University of Wisconsin-Madison, Madison, WI, USA, jjiang1@wisc.edu

N. Jiang Department of Horticulture, Michigan State University, Michigan, USA, jiangn@msu.edu

Eduard Kejnovsky Laboratory of Plant Developmental Genetics, Institute of Biophysics, ASCR, Brno, Czech Republic, kejnovsk@ibp.cz

T.-F. Lee Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA, tzuufen@dbi.udel.edu

T. H. Lee Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA, alfalfa@gmail.com

P.Li Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA, li@dbi.udel.edu

S.-L. Liu Department of Botany, University of British Columbia, Vancouver, Canada, shaolunliu@gmail.com

B.C. Meyers Delaware Biotechnology Institute, University of Delaware, Newark, DE, USA, meyers@dbi.udel.edu

Jeffrey P. Mower Center for Plant Science Innovation, E128 Beadle Center, University of Nebraska Lincoln, Lincoln, NE, USA, jmower2@unl.edu

Gonzalo Nieto-Feliner Real Jardin Botnico, CSIC, Madrid, Spain, nieto@rjb.csic.es

S. Nuthikattu Plant Cellular and Molecular Biology, The Ohio State University, 570 Aronoff Laboratory, Columbus, OH, USA, nuthikattu.1@osu.edu

Andrew H. Paterson Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA, paterson@dogwood.botany.uga.edu

C.L. Richards Department of Integrative Biology, University of South Florida, Tampa, FL, USA, christinalrichards@gmail.com

K. Riha GMI – Gregor Mendel Institute of Molecular Plant Biology GmbH, Vienna, Austria, karel.riha@gmi.oeaw.ac.at

J.A. Rossello Jardín Botnico, Universidad de Valencia, Valencia and Marimurtra Botanical Garden, Carl Faust Foundation, Blanes, Spain, rossello@uv.es

M. Siomos Department of Biochemistry and Cell Biology, University of Vienna, Vienna, Austria, maria.siomos@univie.ac.at

D.B. Sloan Department of Biology, University of Virginia, Charlottesville, VA, USA, dbs4a@virginia.edu

K. Slotkin Plant Cellular and Molecular Biology, The Ohio State University, 570 Aronoff Laboratory, Columbus, OH, USA, slotkin.2@osu.edu

P. Smarda Department of Botany and Zoology, Masaryk University, Brno, Czech Republic, smardap@sci.muni.cz

S. Subramaniam Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, CA, USA, sshabari@gmail.com

H Tang Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA, bao@uga.edu

K.J.F. Verhoeven Department of Terrestrial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands

X. Wang Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA, wangxy@uga.edu

P.G. Wolf College of Science, Department of Biology, Utah State University, Logan, UT, USA, paul.wolf@usu.edu

X. Zhang Department of Plant Biology, University of Georgia, Athens, GA, USA, xiaoyu@plantbio.uga.edu

The First Ten Years of Plant Genome Sequencing and Prospects for the Next Decade 1

Lex E. Flagel and Benjamin K. Blackman

Contents

1.1 Introduction

The genome of *Arabidopsis thaliana*, the first completed plant genome sequence, was published in December of 2000 (The Arabidopsis Genome Initiative 2000). This event marked the beginning of the plant genomics era. Over the next 10 years, there has been swift and striking progress in the field of plant genomics. An outpouring of effort coupled with technological advances has made it possible to sequence, assemble, and analyze the genomes of many additional plant species, including several genomes far larger and more complex than *Arabidopsis*. These changes have ushered in comparative plant genomics, the study of relationships between genomes of different species. Comparative genomic analysis has proven particularly enlightening in revealing recent and ancient events that have impacted the structure and contents of plant genomes; and this field is poised to grow rapidly with the advent of high throughput “Next-Generation” sequencing technologies.

In this chapter we revisit some of the breakthroughs and insights that have emerged from the first decade of plant genomics. We summarize our current understanding of plant genomes, based on more than a dozen published sequences. This is followed by a discussion of the opportunities and challenges that we anticipate over the next decade. We have organized these topics in three chronological phases. In the first phase we focus the on *A. thaliana* genome, which on its own yielded a plethora of discoveries and set the stage for further plant genome sequencing. In the second phase we look beyond *Arabidopsis* to the next wave of plant genomes and the lessons learned from comparing them to *A. thaliana* and to one another. Finally we look forward, into a future where the cost of sequencing a plant genome may be orders of magnitude cheaper than it is today, and we consider the rewards and challenges that this new technological capability will bring.

L.E. Flagel (✉)
Department of Biology, Duke University, 90338, Durham, NC 27708, USA
e-mail: lex.flagel@duke.edu

J.F. Wendel et al. (eds.), *Plant Genome Diversity Volume 1*,
DOI 10.1007/978-3-7091-1130-7_1, © Springer-Verlag Wien 2012

1.2 *Arabidopsis*: The Beginning of the Plant Genomics Era

Scientific studies of *A. thaliana* began in its native Europe in the early 1900s. Seminal research was conducted by Friedrich Laibach and colleagues who developed the first mutants and promoted *A. thaliana* as a tractable species for experimentation (Somerville and Koornneef 2002; Koornneef and Meinke 2010). *Arabidopsis thaliana* possesses a suite of life history traits that make it a convenient study system, including prodigious seed production, small stature, modest growth requirements, cross- and self-pollination compatibility, and a short life cycle. Interest grew steadily among other European researchers who were attracted by the convenience of working with *Arabidopsis* and by stimulating early experiments that demonstrated the power of this new system (Meinke et al. 1998; Somerville and Koornneef 2002; Koornneef and Meinke 2010). By the late 1970s *Arabidopsis* research communities were well-established in the United States and elsewhere, and by the early 1990s—following a string of major discoveries—*A. thaliana* had cemented itself as *the* model organism for many forms of basic plant research (Somerville and Koornneef 2002; Koornneef and Meinke 2010). By this time *Arabidopsis* researchers had developed a rapid genetic transformation system (Feldmann and Marks 1987), and work was well underway toward characterizing thousands of mutants and assessing the natural variation of the species. Simultaneous advances in DNA sequencing technology made whole genome sequencing a reasonable proposition. With an active community and an enviable collection of research tools now in place, it became clear that *A. thaliana*—with its diminutive 125 million base pair genome—was the top plant candidate for whole genome sequencing.

The Arabidopsis Genome Initiative (AGI) was formed in 1996 (Bevan 1997), putting in place the funding, organizational, and intellectual apparatus needed to sequence and analyze the genome. The AGI elected to sequence the accession "Columbia", a fecund inbred line that had been used as the wild type strain in many studies dating back to the 1960s (Koornneef and Meinke 2010). AGI also chose the genome sequencing strategy, the so-called BAC-by-BAC method, which involves recapitulating the linear order of the genome in a series of large genomic fragment clones. Initial reports of the second and fourth chromosome assemblies (Lin et al. 1999; Mayer et al. 1999) were followed by reports of the remaining three chromosomes and a detailed analysis of the complete genome in December of 2000 (The Arabidopsis Genome Initiative 2000; the publications of first, third and fifth chromosomes referenced therein). At the time of publication *A. thaliana* was the third multicellular eukaryote to have its entire genome sequenced, following the nematode (*Caenorhabditis elegans*) and fruit fly (*Drosophila melanogaster*).

The genome assembly consisted of approximately 115.4 million nucleotides. The estimated genome size of *A. thaliana* is approximately 125 million base pairs, however, indicating that approximately 10 million base pairs remained unassembled (The Arabidopsis Genome Initiative 2000). A substantial portion of this unassembled sequence likely originates from highly repetitive areas such as centromeres, and consequently the finished assembly is arranged in ten linear molecules, each representing a chromosome arm from one of *A. thaliana*'s five chromosomes. Now, even a full decade later, much of this unassembled sequence is still unaccounted for, and may remain so for some time to come. Despite this limitation, the assembled portion of the *Arabidopsis* genome would prove to be invaluable to the plant research community.

It is difficult to overstate the novelty and transformative impact of the AGI's findings. Their analyses offered numerous insights, many of which may now seem almost self-evident given the rapid pace of plant genomics research. With the benefit of hindsight, we must consider the weight of these discoveries and basic descriptions at a time when there was little knowledge regarding plant genome architecture. For example, the basic genome organization offered interesting perspectives. Prior to the availability of the *Arabidopsis* genome, cytogeneticists had determined that plant genomes had two well-defined domains, heterochromatin and euchromatin. Heterochromatin consists of densely packed DNA, whereas euchromatin is loosely packed, and both are easily distinguished by chromosome staining techniques. This cytogenetic work suggested that genes would largely be found in euchromatin, because regulatory machinery could access them within the loosely packed structure. In *Arabidopsis*, genes were largely found in regions between the centromere and telomere that closely matched cytogenetically identified euchromatin (Lin et al. 1999; Mayer et al. 1999; The Arabidopsis Genome Initiative 2000). Moreover, transposable elements showed the opposite pattern, being most abundant near the heterochromatic centromeres and rare in gene-rich euchromatin. Despite being an anticipated result, this finding confirmed the general layout of a plant genome and provided a link between the cytogenetic and physical landscapes of the genome. Only a decade ago the aforementioned plant chromosomal organization was an unconfirmed hypothesis; the *Arabidopsis* genome provided the first concrete support.

Additional major insights came from cataloging gene content. As previously mentioned, *A. thaliana* has a small genome, in fact one of the smaller genomes known in plants. It was initially reported that approximately 25,500 genes reside within this relatively small genome (The Arabidopsis

Genome Initiative 2000). Since that time, this number has been revised slightly upward to approximately 27,400 (TAIR 9 genome release; www.arabidopsis.org), reflecting improvements in gene discovery and annotation. Nonetheless, given its small genome size, the *A. thaliana* genome is densely populated with genes, having approximately one gene every 4.5 kilobases (kb) on average (The Arabidopsis Genome Initiative 2000). A surprising amount of insight derives from this simple finding. For one, the total number of genes in *A. thaliana* is greater than in some eukaryotes (*D. melanogaster* and *Saccharomyces cerevisiae*, for example), and comparable to others (for example, *C. elegans* and several mammals). Plants are sessile and must adapt to their local environment, while at the same time maintaining all the necessary equipment needed to generate their own energy and synthesize scores of complex biomolecules. In light of these demands, it is impressive that *A. thaliana* accomplishes these feats with such a modest number of genes. A second implication of this finding is that plants with much larger genomes than *A. thaliana* must either have many more genes or considerably lower gene density. As we know now, the answer appears to be primarily the latter alternative, as no diploid plant genome sequenced to date has more than approximately 60,000 genes (Velasco et al. 2010) despite haploid genome sizes more than an order of magnitude greater than *A. thaliana* (Fig. 1.1).

Beyond simply counting genes, the initial *Arabidopsis* genome analysis focused on the functional characteristics and evolutionary relationships of these genes. By comparing the gene content of *A. thaliana* to other major lineages available at the time (bacteria: *Haemophilus influenzae*, fungi: *S. cerevisiae*, animals: *D. melanogaster* and *C. elegans*), the

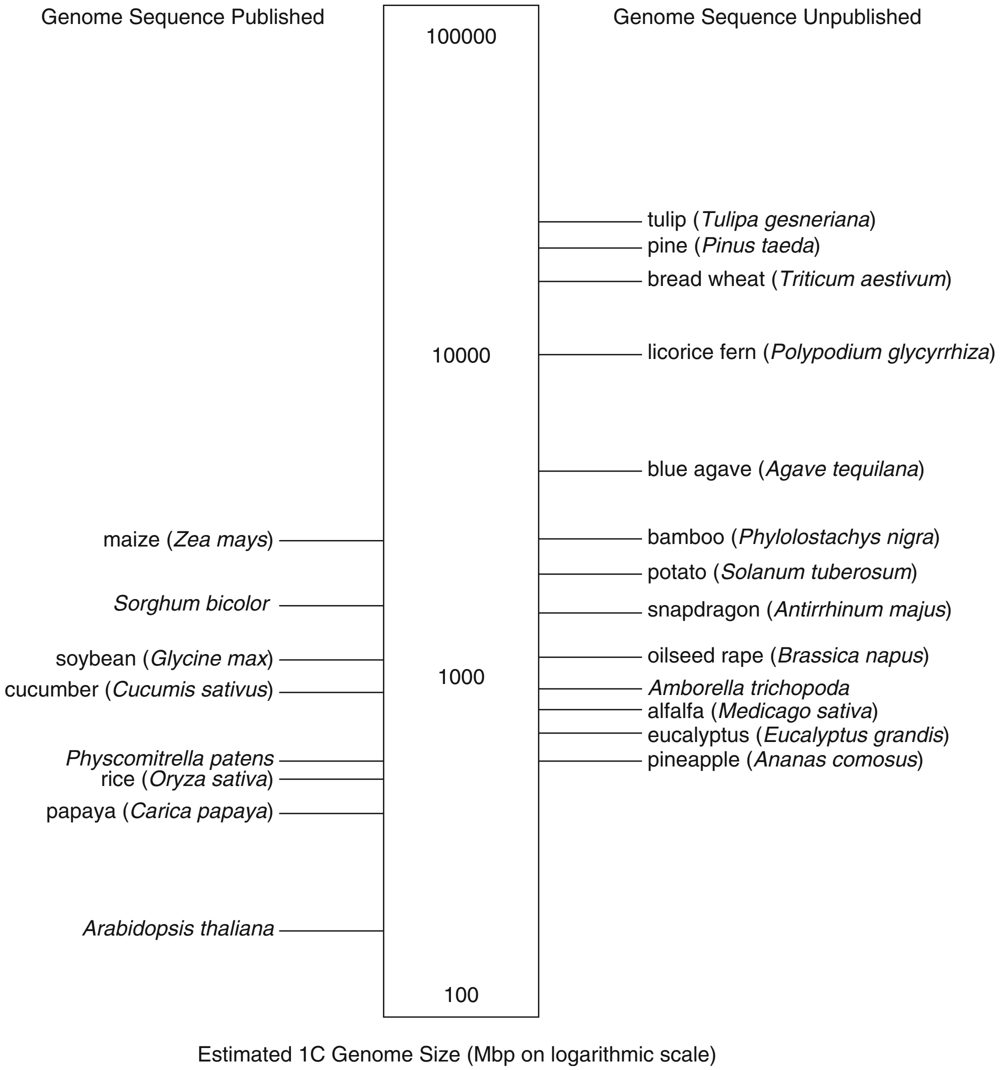

Fig. 1.1 Comparison of plant genome sizes for selected species, including some with (*left side*) and some without (*right side*) sequenced genomes. All plant genome size estimates were taken from the Kew Plant C-values Database (Bennett and Leitch 2010), and are plotted in millions of base pairs (Mbp) on a logarithmic scale

AGI concluded that about 150 major protein families were unique to *A. thaliana* (The Arabidopsis Genome Initiative 2000). At first blush this seems surprisingly low. The divergence of the plant lineage from an ancestral eukaryote is estimated to have occurred 1.5 billion years ago (Yoon et al. 2004), leaving a very long time period for plants to invent new kinds of genes. Two factors likely explain this result. First, there are a large number of basic processes conserved among most organisms—signaling, transcription, genomic maintenance, and metabolic functions are good examples—and these processes involve conserved genes and gene families that encode the core metabolic functions of a living cell. For example, a collection of orthologous eukaryotic proteins show that *A. thaliana* shares 3,285 gene families with at least one other anciently-derived lineage, including the animals *C. elegans*, human, and *D. melanogaster*, the fungi *Schizosaccharomyces pombe* and S. *cerevisiae*, and a basal eukaryote *Encephalitozoon cuniculi* (accessed at NCBI's KOG server: www.ncbi.nlm.nih.gov/COG/). A second explanation for the relative rarity of gene families unique to *A. thaliana* is ascertainment bias. It is easier to annotate known gene families in a newly assembled genome than it is to discover unknown genes. In any case, despite more than a billion years of independent evolution and despite all of the outward differences between plants and other eukaryotes, the *A. thaliana* genome demonstrates that these lineages have much in common at the genetic level.

Another major finding revealed by examining the gene families in the *A. thaliana* genome was their incredible size and redundancy. For example, about 41% of *A. thaliana* genes belong to a gene family with five or more members (The Arabidopsis Genome Initiative 2000), a value considerably higher than what is observed in animal species, where singleton genes are abundant (Lockton and Gaut 2005). Since the publication of the genome sequence, reverse genetic analysis of *A. thaliana* gene families has become something of a cottage industry, with research groups around the world using tools such as T-DNA insertion mutant collections, cDNA and tiling microarrays, and genetic transformation to explore the functions of numerous gene families. Amazingly, some of the largest gene families were scarcely detected before the *A. thaliana* genome sequence became available. One example, the pentatricopeptide repeat genes (PPRs)—including approximately 450 genes in *A. thaliana* and believed to play a role in RNA processing in the mitochondria and chloroplasts—were virtually unknown to plant geneticists prior to a careful analysis of the genome (Aubourg et al. 2000; Small and Peeters 2000; Schmitz-Linneweber and Small 2008). Since their discovery, knockouts of some members of the PPR gene family have been shown to cause disparate and severe mutant phenotypes (e.g., Lurin et al. 2004; Cushing et al. 2005), indicating that they did not escape notice for lack of meaningful function. Another example are the S1 self-incompatibility-like proteins (Ride et al. 1999), which account for approximately 80 genes in the *Arabidopsis* genome and are predicted to be involved in signaling via protein–protein interactions. Despite their large sizes, these gene families "came out of the blue" following whole genome sequencing. These findings speak to the importance of whole genome sequencing in gene discovery and also underscore the enormous role that the *Arabidopsis* genome played in this regard.

The expansion of *A. thaliana* gene families can be largely attributed to duplicated chromosomal segments, many of which are predicted to have arisen via whole genome duplication (polyploidy) (The Arabidopsis Genome Initiative 2000; Vision et al. 2000; Tang et al. 2008a, b). The initial analysis of the *A. thaliana* genome identified 24 large duplications, which in total make up approximately 60% of the genome (The Arabidopsis Genome Initiative 2000). Contemporary *A. thaliana* is a diploid; thus these polyploidy events, apparent from their enduring gene duplications, must have occurred in a distant ancestor (a condition termed *paleopolyploidy*), with the corollary that the genome has since mutationally eroded back to diploidy. The AGI concluded that the duplications were likely the result of an ancient tetraploidy event, as they found evidence for only two pairs of duplicates. As we will see in the next section, comparative genomics has greatly enriched our understanding of paleopolyploidy in plants, and we now have convincing evidence that *Arabidopsis* has experienced multiple rounds of whole genome duplication including an ancient hexaploidy (Ming et al. 2008; Tang et al. 2008a, b).

Beyond sequencing and assembling the genome of the Columbia accession, AGI also partially shotgun sequenced another *A. thaliana* accession, Landsberg *erecta* (The Arabidopsis Genome Initiative 2000). With both genome sequences in hand, these accessions were compared to assess basic parameters of diversity and polymorphism within the species. AGI found that Columbia and Landsberg *erecta* have a single nucleotide polymorphism (SNP) every 3.3 kb, on average, in addition to approximately 14,500 insertion/deletion (indel) polymorphisms scattered throughout the genome. These polymorphisms were found within genes and in intergenic regions. After combining SNPs and indels, the AGI found that 7% of exons had a polymorphism, including numerous indels and amino acid changing substitutions, providing a sizable number of coding region differences that distinguish the two accessions. Finally, they found evidence for more significant structural changes, including transposition events and gene rearrangements. These data were the first assessment of genome-wide polymorphism in a plant species, and ushered in the era of plant *population genomics* (population genetics extended to a genome-wide scope), now a major thrust of plant genome research.

One final virtue of the *Arabidopsis* genome project that merits attention is the path it blazed for future plant genome

Table 1.1 Published plant genomes

	Species	Citation
1	*Arabidopsis thaliana*	The Arabidopsis Genome Initiative (2000)
2	Rice (*Oryza* spp.)	Goff et al. (2002) and Yu et al. (2002)
3	Poplar (*Populus trichocarpa*)	Tuskan et al. (2006)
4	Grape (*Vitis vinifera*)	Jaillon et al. (2007)
5	Papaya *(Carica papaya)*	Ming et al. (2008)
6	*Physcomitrella patens*	Rensing et al. (2008)
7	Maize (*Zea mays*)	Schnable et al. (2009)
8	Sorghum (*Sorghum bicolor*)	Paterson et al. (2009)
9	Apple (*Malus* × *domestica*)	Velasco et al. (2010)
10	*Brachypodium distachyon*	International Brachypodium Initiative (2010)
11	Castor bean (*Ricinus communis*)	Chan et al. (2010)
12	Cucumber (*Cucumis sativus*)	Huang et al. (2009)
13	Soybean (*Glycine max*)	Schmutz et al. (2010)
14	Cacao (*Theobroma cacao*)	Argout et al. (2011)
15	Strawberry (*Fragaria vesca*)	Shulaev et al. (2011)

sequencing efforts in terms of its community structure and public accessibility. At the time of publication there was little precedent for funding, sequencing, and data access for genome sequencing projects. In the late 1990s and early 2000s many differing ideologies and funding models were in circulation, ranging from full public funding with open data access to private funding and data access fees. The *Arabidopsis* genome project presented a model of collaboration that has guided many subsequent plant genome initiatives. It was organized and paid for through a large, multinational collaboration between academic, governmental, and corporate interests, each sharing data and contributing expertise. The genome sequence and annotation were released to the public without access fees. Furthermore, the *Arabidopsis* community organized a web portal called The *Arabidopsis* Information Database (TAIR; www.arabidopsis.org) shortly following publication. This site continues to serve as a primary hub for *Arabidopsis* information, allowing access to sequence data, a genome browser, gene annotations, and ordering information for *Arabidopsis* germplasm. By fostering interaction, sharing, and the foresight to freely distribute genomic information through a centralized hub, the AGI can be credited with putting forth an effective model that has been built upon by many subsequent plant genome sequencing projects.

1.3 The Next Series of Plant Genomes and the Beginning of Comparative Plant Genomics

Arabidopsis opened the floodgates for plant genome sequencing. To our knowledge, the primary descriptions of 15 plant genomes have been published at the time of this writing. Table 1.1 lists these species along with their year of publication and citation, while Fig. 1.2 shows their hypothesized phylogenetic relationships. Beyond these 15 plant species, there are dozens more with genome projects underway, at stages ranging from early planning to near-completion. Indeed, 5 of the 15 published plant genomes in Table 1.1 were released in 2010, and we can expect 2011 and beyond to be equally productive given the large number of ongoing sequencing projects.

As more plant genomes become available, the possibility emerges for functional and evolutionary comparisons among species. Many interesting developments have emerged from this comparative approach to plant genomics. In this section we highlight three key findings—conserved gene order, ancestral polyploidy, and conserved gene content—each with relevance to the future of plant genome sequencing. For in-depth coverage of these topics, as well as other topics in comparative plant genomics, see other chapters in this volume.

Conserved gene order within and between various plant lineages is a major revelation that has emerged over the last several decades. The first hints came from comparing marker order among genetic maps of related species in clades such as the grasses and the Solanaceae (Moore et al. 1995; Gale and Devos 1998). These comparative mapping efforts showed that marker order was often conserved, though in many cases the species being compared no longer shared orthologous chromosomes because of rearrangements. Thus, even over significant evolutionary timescales—encompassing chromosome structural modifications—an ancestral gene order is found between many contemporary plant species. This shared gene order is termed *synteny,* and it is useful to establish as it helps assign orthology and paralogy, the first step in describing evolutionary relationships between genomic regions. Comparative mapping, however, has a limited resolution that

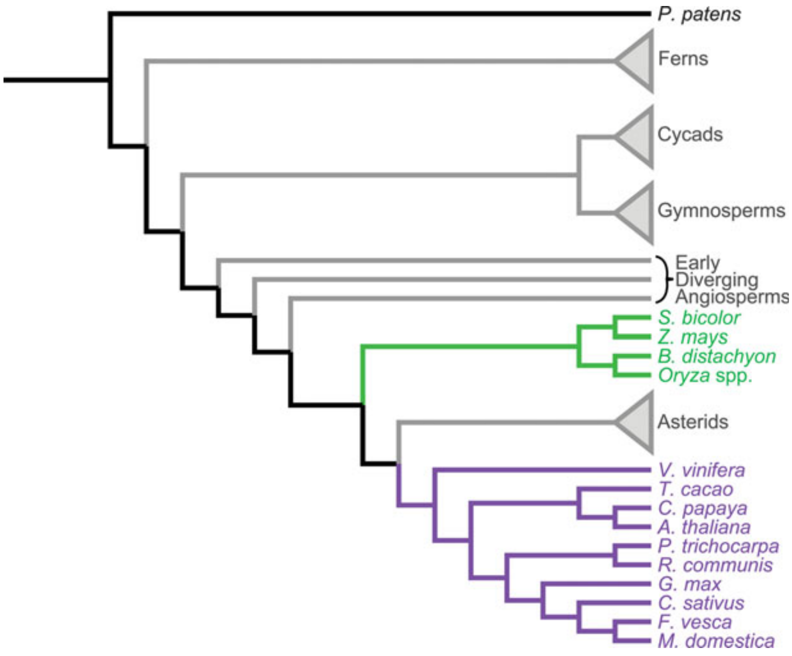

Fig. 1.2 Phylogeny of sequenced land plant genomes. The moss *P. patens* is listed in *black*; all monocots are in *green*; and rosids are in *purple*. Several important lineages lacking a published genome sequence are listed in *grey*. Phylogenetic relationships follow the AGP III system (The Angiosperm Phylogeny Group 2009)

does not permit assessment of the extent to which gene order is conserved at the nucleotide level. Fortunately, following a proliferation of plant genome sequences, we now have a wealth of fine-scale synteny analyses (see the chapter by Paterson et al. (2012, this volume)). The first genomic search for synteny came in the wake of the publication of the second plant genome sequence, rice (*Oryza sativa* spp.) (Goff et al. 2002; Yu et al. 2002), which could be compared to *A. thaliana*. Rice and *Arabidopsis* diverged approximately 200 million years ago—a considerable length of time for the accumulation of chromosomal gains, losses, and rearrangements—and some pre-genomic investigations predicted that little synteny would remain (Devos et al. 1999). This prediction was corroborated by the rice genome sequence, where it was determined that only modest stretches of synteny remained between rice and *A. thaliana* (Goff et al. 2002); the longest stretch contained 119 *Arabidopsis* proteins. Interestingly, among the identified syntenic segments, rice frequently showed a one-to-many relationship with *A. thaliana*, indicating that whole genome duplications in *Arabidopsis* occurred subsequent to its divergence from a common ancestor with rice (Goff et al. 2002). This limited synteny cast doubt over the feasibility of constructing a robust ancestral gene order among the monocots (rice) and eudicots (*Arabidopsis*) and contrasted with previous findings, as noted above, of strong evidence for synteny at closer phylogenetic distances. With the release of several genome sequences more closely related to *Arabidopsis* and rice (Table 1.1 and Fig. 1.2), intra-rosid and intra-monocot searches for synteny could be performed. Within each clade, these comparisons were typically updated with each new genome release, increasing in turn the power and sophistication of the analyses. Looking back, it is apparent that every additional genome sequenced has sharpened our still slightly hazy picture of gene order and its evolutionary dynamics across plants. We need not elaborate this story here, as this has been well covered in the chapter by Paterson et al. (2012, this volume). We will instead jump ahead to highlight pertinent aspects of contemporary knowledge of synteny within the rosids.

At present the rosids—featuring ten published genomes (Fig. 1.2)—offer the most complete view of genomic synteny. The evolutionary history stored within these genomes is a convoluted one. This should not be surprising, in that the rosids include approximately 75,000 species in 17 diverse orders and are found in great abundance in virtually all parts of the world. Moreover, rosid chromosome counts and genome sizes range over at least an order of magnitude (Goldblatt and Johnson 1979; Bennett and Leitch 2010), and concomitant with this structural divergence, gene order also has been greatly disrupted. *Arabidopsis thaliana* in particular has had a wonderfully convoluted genomic history. It is the product of at least three detectable rounds of polyploidy, the last two occurring after its relatively recent divergence from papaya (*Carica papaya*) (Tang et al. 2008a), yet despite multiple genome duplication events, it also has suffered enormous genomic downsizing, both in terms of DNA content and chromosome number. The apparent conflict between this duplication rich history and modest contemporary genome size can only be remedied by invoking massive amounts of genomic loss. Remarkably, *Arabidopsis* still shares tracts of clear synteny with its relatives, and with substantial cutting and pasting its gene order can be projected onto the gene order of other rosid species (Jaillon et al. 2007;

Ming et al. 2008; Tang et al. 2008a). Like *A. thaliana,* poplar (*Populus trichocarpa*) also appears to have a recent history of whole genome duplication, with a single polyploidy event occurring within the salicoid lineage (Tuskan et al. 2006). In contrast to *Arabidopsis* and poplar, the genomes of grape (*Vitis vinifera*) and papaya have had a more static duplication history, showing only a widely shared hexaploidy event embedded at a deep and poorly resolved position within the angiosperm phylogeny (Ming et al. 2008; Tang et al. 2008a, b). Together these findings highlight several intriguing facts. First, there is a predicted hexaploidy event deep within the eudicots, and this hexaploidy may be correlated with the rapid expansion and radiation of these species (De Bodt et al. 2005). Second, through processes that are still poorly understood, some polyploid plant genomes (e.g., *Arabidopsis* and poplar) break down over time and return to diploidy (Wolfe 2001). Finally, some species—like grape and papaya—appear to have been unduplicated for an extensive period of time. These findings bring to light fundamental mysteries about synteny and plant genome structure. For example, how important was polyploidy in driving plant diversification? Also, why do some lineages appear to have dynamic and punctuated genome content evolution while other lineages appear to have maintained a fairly static genomic architecture for millions of years? Answers to these vexing questions are likely to emerge from multiple avenues of investigation, including a continuation and extension of comparative investigations of plant genomes.

Plant researchers have long sought to understand the genetics underlying many of the biological features that are specific to plants. The age of comparative plant genomics has cast a bright light on this area by making it possible to catalog and compare diverse plant gene repertoires. As mentioned earlier, many genes and gene families perform essential tasks and are conserved among living organisms. This is also true within plants; both generic and plant-specific gene families are broadly conserved. Notably, gene structure is also broadly conserved due to low rates of intron gain and loss. For instance, only approximately 5% of genes differ in intron content between *A. thaliana* and rice (Roy and Penny 2007), and exon–intron structure is also highly conserved among poplar, soybean, and grape (Schmutz et al. 2010). However, because of gene duplication and loss, the *size* of conserved genes families varies widely between plant species (Velasco et al. 2007; Rensing et al. 2008). These observations have led some to suggest that amplification and reduction in gene family size could be a more important contributor to evolution and diversification in plants than is the production of novel genes (Flagel and Wendel 2009). In light of this viewpoint, understanding the evolutionary history of plant gene family expansion and contraction can be seen as a tool for elucidating functional evolution.

Early diverging plant lineages—such as the green algae, bryophytes, and lycophytes—offer an enlightening perspective on the dynamic nature of plant gene family size. When compared to angiosperms, these early diverging lineages reveal extensive gene family conservation. For example, the moss *Physcomitrella patens*, which diverged from the angiosperms approximately 450 million years ago, shares 5,809 gene families with *Arabidopsis*, a tally that accounts for approximately 69% of the genes in the *A. thaliana* genome (data extracted from Phytozome version 5.0; www.phytozome.org). In Fig. 1.3 we compare gene family size between some early diverging plants, monocots, and rosids. These comparisons reveal several interesting trends. The green algae *Chlamydomonas reinhardtii* and the lycophyte *Selaginella moellendorffii*—both early diverging plant lineages—generally have fewer genes per shared gene family when compared to angiosperm species (Fig. 1.3). *Physcomitrella patens*, on the other hand, has many conserved gene families that are larger than those found in the angiosperms (Fig. 1.3), which is consistent with the fact that *P. patens* has approximately the same number of predicted genes as *Arabidopsis* (Rensing et al. 2008), while *C. reinhardtii* and *S. moellendorffii* have fewer. Finally, within the angiosperms the grasses rice and maize (*Zea mays*)—which both have a large number of genes (Goff et al. 2002; Yu et al. 2002; Schnable et al. 2009)—also tend to have larger conserved gene families when compared to the rosids *Arabidopsis* and papaya. Taken as a whole these patterns indicate that increases in gene content often result from gene family expansion, rather than novel gene creation.

By looking at the types of genes that have expanded or contracted in each lineage, we may be able to gain insights into the selective pressures plant genomes face. For example, following recent genome duplications in *A. thaliana* and maize lineages there is strong evidence for biased retention of transcription factors (Blanc and Wolfe 2004; Seoighe and Gehring 2004; Schnable et al. 2009). Moreover, in *A. thaliana*, genes involved in basic enzymatic processes and DNA repair appear to have been preferentially *lost* following duplication (Blanc and Wolfe 2004; Seoighe and Gehring 2004). Notably, different conclusions were reached regarding the functional categories preferentially retained in duplicate subsequent to polyploidy events in the Compositae (Barker et al. 2008). Here, transcription factors are underrepresented and genes associated with structural or cellular organization are overrepresented. Finally, a comparison of *A. thaliana*, poplar, rice, and moss revealed that when paralogs arise by tandem duplication rather than polyploidy, genes responsive to abiotic and biotic environmental stimuli are preferentially preserved over evolutionary time (Hanada et al. 2008).

These observations, derived from distantly related species, give an indication that not all duplicated genes have the

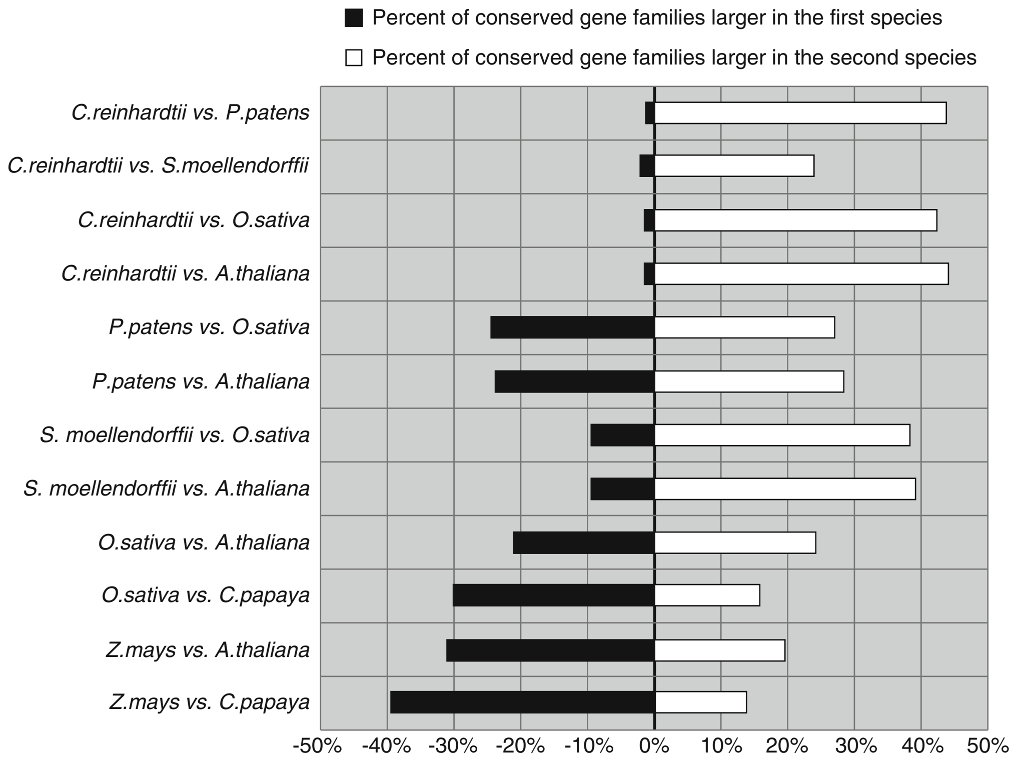

Fig. 1.3 Gene family size comparison between early diverging plant species and angiosperms. Early diverging species include the green alga *Chlamydomonas reinhardtii*, the bryophyte *Physcomitrella patens*, and the lycophyte *Selaginella moellendorffii*. Angiosperms are represented by the monocots *Oryza sativa* and *Zea mays*, and the eudicots *Arabidopsis thaliana* and *Carica papaya*. Select pairwise comparisons of these species are presented, showing the percent of conserved gene families with more genes in the first species (*black*) versus the second species (*white*). Families of the same size in both species are omitted from this tally. All data extracted from Phytozome version 5.0 gene family database (www.phytozome.org)

same probability of retention following duplication. One emerging perspective is that gene family expansion and contraction proceeds in a non-random fashion, and can have secondary effects on interacting genes (Paterson et al. 2010). These patterns may emerge because plants are constantly subjected to new biotic and abiotic stresses, and retaining genes that recognize and react to these agents is adaptively favorable, though neutral processes such as subfunctionalization (the division of ancestral functions among duplicated genes) may also contribute. Because nearly all genes and gene families are embedded within genetic networks—be they regulatory or metabolic—their expansion and contraction can have effects on their interacting partners. Again, looking at the most recent genome duplication in *Arabidopsis* lineage, there is evidence that certain suites of interacting genes are preferentially retained, while their paralogous counterparts are preferentially lost (Thomas et al. 2006). Thus, there is a need to consider gene family expansion and contraction in a network-informed context. Over evolutionary time-scales, these processes can greatly alter gene content, giving rise to highly skewed gene family sizes, and ultimately shaping the entire composition of the genome.

The availability of many plant genome sequences has given rise to an era of comparative plant genomics. By comparing the structure and contents of various species in a phylogenetic context we are beginning to discover the evolutionary forces that have shaped modern plant genomes. This work has shown that plants possess a conserved gene repertoire, yet each species experiences significant gains and losses of genetic material. This occurs in the form of recent and ancient genome duplication, as well as lineage-specific gene family expansion and contraction. Future work will continue to shed light on these unique aspects of genomic evolution in plants.

1.4 Moving Forward: The Second Decade of Plant Genomics

After considering the great strides plant genomics has made in the last decade, we now take the opportunity to imagine the potential triumphs and challenges that lie ahead. But, before we delve into these topics, it is worth considering the rationale for continued plant genome sequencing. As documented in the previous section, full genome sequences are available or in development for several plant research models and many of the world's major crop species. Notwithstanding these many accomplishments, existing genome descriptions remain in some respects incomplete. That is, every genome assembly contains missing or inaccurate information, and this incompleteness is exacerbated by the fact that all available genome sequences, even *A. thaliana*, contain thousands of annotated genes that lack an experimentally determined function. Thus, it could be argued that we have already surpassed our capacity to study the genomic information we have in hand, or to put it another way, the first decade of plant genomics has created an information reservoir sufficient to fuel decades of follow-up research. Based on this, an argument could be made that new resources might be more optimally diverted toward a more complete analysis of existing genomes. In contrast to this perspective, here we present the case that continued investment in plant genome sequencing is warranted. Our primary reasoning is based on the recognition that genome sequencing is both a tool for new discovery and a means of achieving functional insight. In this section we discuss the power of genome sequencing in both of these capacities, demonstrating how genome sequencing can be used to reveal the functions of known genes, in addition to enabling new and unexpected discoveries.

Genome sequencing itself is merely a starting point in the analysis of the molecular and cellular function of genes. Immediately after release, many useful tools can be developed to open up an array of new research questions. Classic examples include sequence-based marker development, gene-targeted mutational forward genetics, and microarray platform design for gene expression and copy number variation studies (e.g. Schmid et al. 2005; Borevitz et al. 2007; Springer et al. 2009). In addition, sophisticated tools can be developed to access small RNAs and probe epigenetic modifications (e.g. Zhang et al. 2006; Lister et al. 2009). All of these tools have consistently led to new discoveries, and they will continue to be a primary impetus for further genome sequencing. Moreover, the benefits of plant genome sequencing continue to emerge with the addition of new technologies. For example, with a high-quality reference genome, additional individuals can be cost effectively *resequenced* (generating and aligning short Next-Generation sequence reads from a new genotype to a reference genome to produce a novel draft genome sequence). Genomic information from resequenced individuals can in turn be used understand genetic diversity within species. We are now beginning to see this concept in action and once again *Arabidopsis* has taken the lead with the 1001 Genomes Project, which aims to sequence and disseminate genome-wide diversity data from 1001 *Arabidopsis* accessions (Ossowski et al. 2008). Today, a researcher interested in an *Arabidopsis* gene can scan polymorphism data at the 1001 Genomes Project website (www.1001genomes.org), and identify accessions with interesting polymorphisms without ever entering a laboratory. As expected, this resource is quickly replacing the time consuming process of cloning and sequencing a gene of interest from various *Arabidopsis* accessions. A major lesson has been learned from the initial genome sequencing projects: one genome representing an entire species is simply not sufficient for many research questions. More powerful forms of analysis, utilizing principles from population genetics, can be performed when multiple genome sequences are available for a species. Along with the continued development of classical genomic tools, projects similar to the *Arabidopsis* 1001 Genomes Project have been initiated for a few major plant research models and crops (Huang et al. 2010; Lai et al. 2010; Lam et al. 2010). With the steadily falling price of DNA sequencing, we expect this to be an area of great growth in the next decade.

Additional information can be extracted from genomic data once it is put in an evolutionary context because genomic features can be analyzed in a phylogenetic manner. Consequently, the accumulation of new genomes enriches the value of existing genome sequences. Notably, plants have possibly the best-resolved phylogeny of any major eukaryotic lineage (Savolainen and Chase 2003), which provides a powerful framework for evolutionary comparison. The first available plant genome sequences came from distantly related species, making them only appropriate for broad evolutionary comparisons. As discussed in the previous section and in additional chapters throughout this volume, these broad comparisons have proved fruitful. The next phase of comparative genomics will likely harness the evolutionary information found in closely related species, perhaps at the relatively shallow taxonomic depths of populations or genera. These analyses will benefit greatly from the strong evolutionary signal that is maintained at relatively close phylogenetic proximity. With multiple genome sequences available at the genus level we may someday be able to study genome evolution with sufficient precision to ascertain the exact spectrum and perhaps sequence of molecular evolutionary events that have shaped chromosome organization or gene family composition. As an example of the potential of fine-scale evolutionary analysis, various monocot genomes were utilized to demonstrate a history of chromosome fusions within the *Brachypodium distachyon* genome (International Brachypodium Initiative 2010). This example shows that some plant lineages are approaching the critical mass of fully sequenced species needed to make precise evolutionary comparisons (notably the Brassicaceae and Poaceae; Fig. 1.2). In the future, expanded plant genome sequencing efforts will only continue to add more power to this approach, further enabling our ability to discover new aspects of plant genome evolution. Similarly, combining new computational tools with the evolutionary signal from additional plant genomes of intermediate phylogenetic distance will substantially enhance our power to functionally annotate the coding and regulatory portions of the genome (Wang et al. 2009; Picot et al. 2010). For many species we are, or will soon be able, to annotate and analyzed their genomes in a phylogenetic framework. In the future this approach will dramatically benefit the quality and utility of plant genome sequences.

1.5 Challenges That Lie Ahead

Though the plant genomics community has made impressive progress in the decade following the release of the *Arabidopsis* genome, many critical plant species remain to be sequenced (some can be found in Figs. 1.1, 1.2). Among the species without a sequenced genome are important crop plants, including, for example, coffee (*Coffea arabica*), common bean *(Phaseolus vulgaris*), cotton (*Gossypium hirsutum*), potato (*Solanum tuberosum*), and sugarcane (*Saccharum* spp.). Other species lacking genome sequences are important research models such as snapdragon (*Antirrhinum majus*) and tobacco (*Nicotiana tabaccum*; both a crop and research model). Finally, there are large clades in the angiosperm tree of life that have yet to be included in the tabulation of sequenced genomes, including early diverging angiosperms, asterids, and several anciently diverged lineages such as the ferns, cycads, and gymnosperms (Fig. 1.2; note some of the aforementioned species have ongoing genome sequencing projects). Most striking among these might be the asterids, one of the two major eudicot clades, which at present have no representative members with a published genome sequence (Fig. 1.2). The asterids include over 60,000 species and contain numerous crop species, including coffee, potato, sunflower (*Helianthus annuus*), and olive (*Olea europaea*), to name just a few. Interestingly, the other comparably large and diverse eudicot clade, the rosids, makes up the majority of published plant genome sequences (Fig. 1.2). This contrast will likely soon end, as draft asterid genome sequences are in the finishing stages for the monkey flower (*Mimulus guttatus*) and tomato (*Solanum lycopersicum*) (pers. comm. T.J. Vision). Over the next decade many of the pockets of uninvestigated plant diversity will likely be explored, because there are important species in these clades and because understanding the full biodiversity of plant life is a major goal of many research organizations around the world.

Possibly the greatest future challenge in plant genome sequencing will be tackling the massive genomes found among some of the world's most important plants. Figure 1.1 displays the genome sizes of a number of published plant genome sequences, compared to a range of unpublished plant species selected for their economic, cultural, or research significance. Clearly we have only completed genome sequencing projects at the bottom of the range of plant genome sizes. Some exceptionally important plants have very large genome sizes, and sequencing these genomes will be a great challenge. As an example, upland cotton (*Gossypium hirsutum*), bread wheat (*Triticum aestivum*), and pines (species in the genus *Pinus*) have genome sizes that are approximately 21, 108, and 138 times the size of *Arabidopsis*, respectively (Bennett and Leitch 2010). Many essential human activities are reliant on these species, and it seems imperative that at some point we will endeavor to sequence their genomes. What we might find in the complex genomes of these species will be quite exciting. We know that, in part, upland cotton and bread wheat have large genome sizes because they are polyploids (cotton is an allotetraploid while bread wheat is an allohexaploid). To our knowledge, no recent polyploid plant genome sequence has been published (though some recent paleopolyploids have, such as maize and soybean (Schnable et al. 2009; Schmutz et al. 2010)). The duplicate nature of polyploid genomes creates several hurdles for accurate assembly. For example, polyploids with divergent genomes (allopolyploids, such as cotton and bread wheat) add complexity as both co-resident genomes will need to be differentiated and assembled separately. Difficulties aside, many critically important plant species are polyploid, creating a strong motivation to surmount these challenges. Notably, despite their massive genome size, most pines are thought to be diploid, and what makes their genome so large is a fascinating mystery. The pine genome could be beset with enormous numbers of transposable elements or replete with segmental duplications. At present, producing a quality pine genome would be a tremendous undertaking, both technically and monetarily, but if the next decade witnesses similar technological improvements and cost reductions as the last, the feasibility of completing a pine genome sequence could be just on the horizon (Neale and Kremer 2011).

The discussion of massive and complex plant genomes raises another interesting question: is it necessary to have a nearly complete genome assembly, or could great advances be stimulated by much less expensive and less technically demanding low-coverage sequencing scans? The answer, of course, depends on how the research community intends to use the information. Many of the first genome sequences originated from model organisms, and the research communities backing these species often required a robust and detailed understanding of genomic structure and content. On the other hand, a low-coverage and incomplete assembly may be all that is needed to aid some forms of plant improvement, like marker-based selection. In the next decade, plant genome sequencing endeavors will move beyond the most utilized model species and the smallest genome sizes, and we may see a concomitant adjustment in the level of genomic detail that is sufficient for the research goals of these communities. Some movement in this direction is already evident from the great abundance of plant expressed sequence tag (EST) projects. EST sequencing is simply the capture and sequencing of mRNA transcripts, an effective way to isolate the expressed portion of the genome. Surely, EST sequencing—like low-coverage genome sequencing—has major limitations; however, for some research questions these low cost alternatives may provide sufficient data. Nevertheless, we anticipate a future where most research communities will seek whole plant genome sequences, because this resource facilitates the greatest range of genomic tools and analyses and offers insights that cannot be matched by low-coverage alternatives. Indeed, in many cases, low-coverage sequences are likely to be used as a stepping-stone toward the eventual goal of producing a complete genome sequence.

Current and future plant genome sequencing projects are greatly aided by new sequencing technologies, which have caused a precipitous drop in the price per base pair of DNA sequence (Huang et al. 2009; Argout et al. 2011; Shulaev et al. 2011). In a just a matter of weeks, a small group of researchers can now produce billions of nucleotides of sequence for only thousands of US dollars. Though the price of sequencing has fallen dramatically, *assembling* a new plant genome is far from routine or easy. Assembly still requires significant infrastructure, technological proficiency, and effort. The assembly issue has become even thornier with current Next-Generation sequencing technologies, primarily because these platforms produce short reads (generally between 30 and 500 base pairs in length). For most plant genomes it is unlikely that reads of these lengths will assemble into large contiguous pieces, instead producing a highly fragmented assembly that offers more limited utility. In large part this fragmentation occurs because many plant repeats are significantly longer than these short read lengths and thus cannot be traversed, effectively terminating any non-arbitrary elongation from the point of the repeat onward (Pop and Salzberg 2008). A greater depth of sequence coverage—the area in which Next-Generation technologies excel—cannot alleviate this problem. Instead, alternative library construction and sequencing techniques, often acquired at a far greater cost, must be used to bridge these large repeats. Thus, despite the impressive technological gains of the last few years, producing a high-quality genome remains an expensive and laborious task. One obvious way to overcome this limitation is the development of sequencing technologies that can produce longer reads (for example >1,000 base pairs, approximately the current ceiling for Sanger sequencing technology) in a rapid and cost effective manner. Exceptionally long reads (>10,000 base pairs), could potentially be an even larger advance, as they are likely to span many plant repeats and consequently greatly reduce the complexity of genome assembly. Single-molecule sequencing (Eid et al. 2009), is emerging as a potential front-runner in this area. Single-molecule sequencers that produce read lengths comparable to Sanger sequencing are beginning to appear, though it is too early to predict where the technology might be headed. It does, however, seem likely that over the next decade DNA sequencing devices with tremendous throughput and exceptionally long read lengths will be developed. This will prove to be a breakthrough for plant genome sequencing efforts. It will precipitate new strategies for genome sequencing and assembly, and likely open up the door for plant genome sequencing in species that have largely been neglected by funding agencies.

Finally, to better utilize future plant genome sequences, researchers will need to integrate data between research groups. At present, a research community typically conducts in-house genome assembly and annotation, and stores this information in a species-specific database. So far this model has served each community well, but has also created some impediments to multi-species comparative analyses. The plant community would benefit greatly from unified data annotation terminology and distribution models. For example, in some species probable pseudo-genes are left in the final gene catalog, while in other species they are systematically excluded. If one wishes to compare genes between these species, one must first perform a tedious filtering task. Fortunately, genome database experts have undertaken some of this data integration. A few good examples of well-curated, multi-species, comparative plant genome databases include Phytozome (www.phytozome.org), Plant Genome Database (www.plantgdb.org), and Plant Genome Duplication Database (http://chibba.pgml.uga.edu/duplication/). The research community behind each newly sequenced species will want to rapidly integrate their new sequence into a framework with existing plant genome sequences for comparative purposes. This creates an incentive to adopt a common framework. We anticipate that this motivation will stimulate researchers to gravitate toward packaging their data in a standardized way, which will ultimately benefit the entire research community.

Conclusions

Tremendous resources have been devoted to plant genome sequencing. This outlay has thus far rewarded us with 15 published sequences. Today—one decade into the plant genomics era—comparative plant genomics is beginning to realize its promise, revealing the nature of the evolutionary forces that have shaped the structure and contents of modern plant genomes. In the future, with the advent of improved sequencing technologies, comparative plant genomics will continue to be a major a source of new insights. Existing plant genomes have also greatly benefited plant research as a platform for gaining functional insights. Work along these lines will continue to flourish, including greater within-species genome resequencing, which will prove to be a key method for elucidating the genetic consequences of population level processes. There is cause for continued optimism about future developments in these areas, particularly as we surmount the challenges inherent in sequencing and assembling massive and redundant plant genomes. Similarly, we anticipate continued progress in integrating genomic data from disparate sources, creating new opportunities for comparative discovery. Because of this, we anticipate that the second decade of plant genomics will surpass the first in terms of scientific breakthroughs and advances to our knowledge of plant diversity.

References

Argout X, Salse J, Aury J-M, Guiltinan MJ, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova SN, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Barbosa-Neto JF, Sabot F, Kudrna D, Ammiraju JSS, Schuster SC, Carlson JE, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Berard A, Viot C, Boccara M, Risterucci AM, Guignon V, Sabau X, Axtell MJ, Ma Z, Zhang Y, Brown S, Bourge M, Golser W, Song X, Clement D, Rivallan R, Tahi M, Akaza JM, Pitollat B, Gramacho K, D'Hont A, Brunel D, Infante D, Kebe I, Costet P, Wing R, McCombie WR, Guiderdoni E, Quetier F, Panaud O, Wincker P, Bocs S, Lanaud C (2011) The genome of *Theobroma cacao*. Nat Genet 43:101–108

Aubourg S, Boudet N, Kreis M, Lecharny A (2000) In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. Plant Mol Biol 42:603–613

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Mol Biol Evol 25:2445–2455

Bennett M, Leitch I (2010) Plant DNA C-values database (release 5.0). http://www.kew.org/cvalues/

Bevan M (1997) Objective: the complete sequence of a plant genome. Plant Cell 9:476–478

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell 16:1679–1691

Borevitz JO, Hazen SP, Michael TP, Morris GP, Baxter IR, Hu TT, Chen H, Werner JD, Nordborg M, Salt DE, Kay SA, Chory J, Weigel D, Jones JDG, Ecker JR (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 104:12057–12062

Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G, Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD (2010) Draft genome sequence of the oilseed species *Ricinus communis*. Nat Biotechnol 28:951–956

Cushing DA, Forsthoefel NR, Gestaut DR, Vernon DM (2005) *Arabidopsis emb175* and other *ppr* knockout mutants reveal essential roles for pentatricopeptide repeat (PPR) proteins in plant embryogenesis. Planta 221:424–436

De Bodt S, Maere S, Van de Peer Y (2005) Genome duplication and the origin of angiosperms. Trends Ecol Evol 20:591–597

Devos KM, Beales J, Nagamura Y, Sasaki T (1999) *Arabidopsis*-rice: will collinearity allow gene prediction across the eudicot–monocot divide? Genome Res 9:825–829

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Feldmann KA, Marks MD (1987) *Agrobacterium*-mediated transformation of germinating seeds of *Arabidopsis thaliana*: a non-tissue culture approach. Mol Gen Genet 208:1–9

Flagel LE, Wendel JF (2009) Gene duplication and evolutionary novelty in plants. New Phytol 183:557–564

Gale MD, Devos KM (1998) Plant comparative genetics after 10 years. Science 282:656–659

Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun W-l, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Goldblatt P, Johnson D (1979) Index to plant chromosome numbers. Missouri Botanical Garden, St. Louis

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu S-H (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol 148:993–1003

Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, Ren Y, Zhu H, Li J, Lin K, Jin W, Fei Z, Li G, Staub J, Kilian A, van der Vossen EAG, Wu Y, Guo J, He J, Jia Z, Ren Y, Tian G, Lu Y, Ruan J, Qian W, Wang M, Huang Q, Li B, Xuan Z, Cao J, Asan WuZ, Zhang J, Cai Q, Bai Y, Zhao B, Han Y, Li Y, Li X, Wang S, Shi Q, Liu S, Cho WK, Kim J-Y, Xu Y, Heller-Uszynska K, Miao H, Cheng Z, Zhang S, Wu J, Yang Y, Kang H, Li M, Liang H, Ren X, Shi Z, Wen M, Jian M, Yang H, Zhang G, Yang Z, Chen R, Liu S, Li J, Ma L, Liu H, Zhou Y, Zhao J, Fang X, Li G, Fang L, Li Y, Liu D, Zheng H, Zhang Y, Qin N, Li Z, Yang G, Yang S, Bolund L, Kristiansen K, Zheng H, Li S, Zhang X, Yang H, Wang J, Sun R, Zhang B, Jiang S, Wang J, Du Y, Li S (2009) The genome of the cucumber, *Cucumis sativus* L. Nat Genet 41:1275–1281

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang Q-F, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

International Brachypodium Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature 463:763–768

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Koornneef M, Meinke D (2010) The development of *Arabidopsis* as a model plant. Plant J 61:909–921

Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, Xiang Z, Song W, Ying K, Zhang M, Jiao Y, Ni P, Zhang J, Li D, Guo X, Ye K, Jian M, Wang B, Zheng H, Liang H, Zhang X, Wang S, Chen S, Li J, Fu Y, Springer NM, Yang H, Wang J, Dai J, Schnable PS, Wang J (2010) Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet 42:1027–1030

Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SS-M, Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059

Lin X, Kaul S, Rounsley S, Shea TP, Benito M-I, Town CD, Fujii CY, Mason T, Bowman CL, Barnstead M, Feldblyum TV, Buell CR, Ketchum KA, Lee J, Ronning CM, Koo HL, Moffat KS, Cronin LA, Shen M, Pai G, Van Aken S, Umayam L, Tallon LJ, Gill JE, Adams MD, Carrera AJ, Creasy TH, Goodman HM, Somerville CR, Copenhaver GP, Preuss D, Nierman WC, White O, Eisen JA, Salzberg SL, Fraser CM, Venter JC (1999) Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761–768

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo Q-M, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322

Lockton S, Gaut BS (2005) Plant conserved non-coding sequences and paralogue evolution. Trends Genet 21:60–65

Lurin C, Andres C, Aubourg S, Bellaoui M, Bitton F, Bruyere C, Caboche M, Debast C, Gualberto J, Hoffmann B, Lecharny A, Le Ret M, Martin-Magniette M-L, Mireau H, Peeters N, Renou J-P, Szurek B, Taconnat L, Small I (2004) Genome-wide analysis of *Arabidopsis* pentatricopeptide repeat proteins reveals their essential role in organelle biogenesis. Plant Cell 16:2089–2103

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N, Harris B, Ansorge W, Brandt P, Grivell L, Rieger M, Weichselgartner M, de Simone V, Obermaier B, Mache R, Muller M, Kreis M, Delseny M, Puigdomenech P, Watson M, Schmidtheini T, Reichert B, Portatelle D, Perez-Alonso M, Boutry M, Bancroft I, Vos P, Hoheisel J, Zimmermann W, Wedler H, Ridley P, Langham SA, McCullagh B, Bilham L, Robben J, Van der Schueren J, Grymonprez B, Chuang YJ, Vandenbussche F, Braeken M, Weltjens I, Voet M, Bastiaens I, Aert R, Defoor E, Weitzenegger T, Bothe G, Ramsperger U, Hilbert H, Braun M, Holzer E, Brandt A, Peters S, van Staveren M, Dirkse W, Mooijman P, Lankhorst RK, Rose M, Hauf J, Kotter P, Berneiser S, Hempel S, Feldpausch M, Lamberth S, Van den Daele H, De Keyser A, Buysshaert C, Gielen J, Villarroel R, De Clercq R, Van Montagu M, Rogers J, Cronin A, Quail M, Bray-Allen S, Clark L, Doggett J, Hall S, Kay M, Lennard N, McLay K, Mayes R, Pettett A, Rajandream MA, Lyne M, Benes V, Rechmann S, Borkova D, Blocker H, Scharfe M, Grimm M, Lohnert TH, Dose S, de Haan M, Maarse A, Schafer M, Muller-Auer S, Gabel C, Fuchs M, Fartmann B, Granderath K, Dauner D, Herzl A, Neumann S, Argiriou A, Vitale D, Liguori R, Piravandi E, Massenet O, Quigley F, Clabauld G, Mundlein A, Felber R, Schnabl S, Hiller R, Schmidt W, Lecharny A, Aubourg S, Chefdor F, Cooke R, Berger C, Montfort A, Casacuberta E, Gibbons T, Weber N, Vandenbol M, Bargues M, Terol J, Torres A, Perez-Perez A, Purnelle B, Bent E, Johnson S, Tacon D, Jesse T, Heijnen L, Schwarz S, Scholler P, Heber S, Francs P, Bielke C, Frishman D, Haase D, Lemcke K, Mewes HW, Stocker S, Zaccaria P, Bevan M, Wilson RK, de la Bastide M, Habermann K, Parnell L, Dedhia N, Gnoj L, Schutz K, Huang E, Spiegel L, Sehkon M, Murray J, Sheet P, Cordes M, Abu-Threideh J, Stoneking T, Kalicki J, Graves T, Harmon G, Edwards J, Latreille P, Courtney L, Cloud J, Abbott A, Scott K, Johnson D, Minx P, Bentley D, Fulton B, Miller N, Greco T, Kemp K, Kramer J, Fulton L, Mardis E, Dante M, Pepin K, Hillier L, Nelson J, Spieth J, Ryan E, Andrews S, Geisel C, Layman D, Du H, Ali J, Berghoff A, Jones K, Drone K, Cotton M, Joshu C, Antonoiu B, Zidanic M, Strong C, Sun H, Lamar B, Yordan C, Ma P, Zhong J, Preston R, Vil D, Shekher M, Matero A, Shah R, Swaby IK, O'Shaughnessy A, Rodriguez M, Hoffman J, Till S, Granat S, Shohdy N, Hasegawa A, Hameed A, Lodhi M, Johnson A, Chen E, Marra M, Martienssen R, McCombie WR (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature 402:769–777

Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M (1998) *Arabidopsis thaliana*: a model plant for genome analysis. Science 282:662–682

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KLT, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang M-L, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na J-K, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Perez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo M-C, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, dePamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996

Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution: grasses, line up and form a circle. Curr Biol 5:737–739

Neale DB, Kremer A (2011) Forest tree genomics: growing resources and applications. Nat Rev Genet 12:111–122

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res 18:2024–2033

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob ur R, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Paterson AH, Freeling M, Tang H, Wang X (2010) Insights from the comparison of plant genome sequences. Annu Rev Plant Biol 61:349–372

Paterson A, Wang X, Tang H, Lee TH (2012) Synteny and genomic rearrangements. In: Wendel JF (ed) Plant genome diversity, vol 1, Plant genomes, their residents, and their evolutionary dynamics. Springer-Verlag, Wien, New York

Picot E, Krusche P, Tiskin A, Carré I, Ott S (2010) Evolutionary analysis of regulatory sequences (EARS) in plants. Plant J 64:165–176

Pop M, Salzberg SL (2008) Bioinformatics challenges of new sequencing technology. Trends Genet 24:142–149

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud P-F, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin IT, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S-i, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu S-H, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69

Ride JP, Davies EM, Franklin FCH, Marshall DF (1999) Analysis of *Arabidopsis* genome sequence reveals a large new gene family in plants. Plant Mol Biol 39:927–932

Roy SW, Penny D (2007) Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*. Mol Biol Evol 24:171–181

Savolainen V, Chase MW (2003) A decade of progress in plant molecular phylogenetics. Trends Genet 19:717–724

Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet 37:501–506

Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci 13:663–670

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-C, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh C-T, Emrich SJ, Jia Y, Kalyanaraman A, Hsia A-P, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia J-M, Deragon J-M, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. Trends Genet 20:461–464

Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, Jaiswal P, Mockaitis K, Liston A, Mane SP, Burns P, Davis TM, Slovin JP, Bassil N, Hellens RP, Evans C, Harkins T, Kodira C, Desany B, Crasta OR, Jensen RV, Allan AC, Michael TP, Setubal JC, Celton J-M, Rees DJG, Williams KP, Holt SH, Rojas JJR, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA, Troggio M, Viola R, Ashman T-L, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Bryant DW, Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Girona EL, Zdepski A, Wang W, Kerstetter RA, Schwab W, Korban SS, Davik J, Monfort A, Denoyes-Rothan B, Arus P, Mittler R, Flinn B, Aharoni A, Bennetzen JL, Salzberg SL, Dickerman AW, Velasco R, Borodovsky M, Veilleux RE, Folta KM (2011) The genome of woodland strawberry (*Fragaria vesca*). Nat Genet 43:109–116

Small ID, Peeters N (2000) The PPR motif—a TPR-related motif prevalent in plant organellar proteins. Trends Biochem Sci 25:45–47

Somerville C, Koornneef M (2002) A fortunate choice: the history of *Arabidopsis* as a model plant. Nat Rev Genet 3:883–889

Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, Iniguez AL, Barbazuk WB, Jeddeloh JA, Nettleton D, Schnable PS (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. PLoS Genet 5:e1000734

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008a) Synteny and collinearity in plant genomes. Science 320:486–488

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008b) Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Res 18:1944–1954

The Angiosperm Phylogeny Group (2009) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121

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

Thomas BC, Pedersen B, Freeling M (2006) Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res 16:934–946

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, dePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé J-C, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai C-J, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, FitzGerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Demattè L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS One 2: e1326

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar SK, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S, Zini E, Eldredge G, Fitzgerald LM, Gutin N, Lanchbury J, Macalma T, Mitchell JT, Reid J, Wardell B, Kodira C, Chen Z, Desany B, Niazi F, Palmer M, Koepke T, Jiwan D, Schaeffer S, Krishnan V, Wu C, Chu VT, King ST, Vick J, Tao Q, Mraz A, Stormo A, Stormo K, Bogden R, Ederle D, Stella A, Vecchietti A, Kater MM, Masiero S, Lasserre P, Lespinasse Y, Allan AC, Bus V, Chagne D, Crowhurst RN, Gleave AP, Lavezzo E, Fawcett JA, Proost S, Rouze P, Sterck L, Toppo S, Lazzari B, Hellens RP, Durel C-E, Gutin A, Bumgarner RE, Gardiner SE, Skolnick M, Egholm M, Van de Peer Y, Salamini F, Viola R (2010) The genome of the domesticated apple (*Malus* × *domestica* Borkh.). Nat Genet 42:833–839

Vision TJ, Brown DG, Tanksley SD (2000) The origins of genomic duplications in *Arabidopsis*. Science 290:2114–2117

Wang X, Haberer G, Mayer K (2009) Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation. BMC Genomics 10:284

Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. Nat Rev Genet 2:333–341

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21:809–818

Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X, Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W, Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J, Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z, Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92

Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SWL, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE, Ecker JR (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell 126:1189–1201

2 Plant Transposable Elements: Biology and Evolution

Eduard Kejnovsky, Jennifer S. Hawkins, and Cédric Feschotte

Contents

2.1 Introduction

Beginning with the pioneering work in the 30s and 40s of Barbara McClintock, R.A. Brink, Rollins Emerson, Marcus Rhoades, and other prominent maize geneticists, transposable elements (TEs) have come to occupy a central position in the study of plant genomes. Not only did McClintock's discovery of the *Activator*/*Dissociation* (*Ac*/*Ds*) system of maize change forever our appreciation of the dynamic nature of chromosomes, her seminal characterization of the regulatory influence of 'controlling elements' (such as *Ac*/*Ds* and later the *Enhancer*/*Suppressor-Mutator* (*En*/*Spm*) system) on adjacent gene expression paved the way for decades of exciting research on the control, both genetic and epigenetic, of gene regulation in plants and other eukaryotes.

It took four decades after McClintock's groundbreaking discoveries and the rise of recombinant DNA technology for the first TEs to be cloned and sequenced in the 1980s. One of the surprises from these early molecular studies was the striking similarity in structure, genetic organization, and even sometimes nucleotide sequence, among the first TEs characterized in maize, snapdragon, *Drosophila* and bacteria (Green 1980; Fedoroff et al. 1983; Levis et al. 1984; Saedler et al. 1984). At that time, and over the next two decades, the biology of TEs was assessed primarily on the basis of the mutations they engendered. Myriad mutant alleles caused by insertions and/or rearrangements of transposons were collected by geneticists in the field, the greenhouse and the fly room, and meticulously analyzed at the molecular level in the lab. Although this era furnished many crucial insights regarding the mechanistic underpinnings and mutagenic capabilities of transposition (for review, Berg and Howe 1989), it yielded little information regarding the abundance and diversity of TEs, much less the long-term evolutionary impact of TE activity.

The advent of large-scale DNA sequencing over the last two decades, combined with advances in functional

E. Kejnovsky (✉)
Institute of Biophysics, ASCR, Kralovopolska 135, 612 00 Brno, Czech Republic
e-mail: kejnovsk@ibp.cz

J.F. Wendel et al. (eds.), *Plant Genome Diversity Volume 1*,
DOI 10.1007/978-3-7091-1130-7_2, © Springer-Verlag Wien 2012

genomics and bioinformatics, has transformed the study of TE biology. This “genomics revolution” has resulted in a greater understanding of the many ways that TEs influence the function and evolution of genes and genomes, and consequently, their host organisms. In particular the genomics era has revealed that, although only a tiny fraction of TEs are transpositionally active, most eukaryotic genomes, and especially plant genomes, are packed with a plethora of seemingly dormant or inactivated TE families (Feschotte et al. 2002). Given the inherent mutagenic potential of active transposition, it should come as no surprise that the majority of these TEs are either defective, fossilized copies or potentially active copies that are restrained by host silencing systems; however, active transposition, as evidenced by instances of mutagenic (yet potentially evolutionarily significant) insertions, has been demonstrated. For example, TEs have been shown to silence or alter expression of genes adjacent to insertion sites, become integrated into functional genes as newly acquired exons (exapted), acquire host gene sequences and insert them into new genomic locations, contribute to chromosomal rearrangements via recombination, epigenetically alter regional methylation patterns, and provide template sequences for RNA interference (Feschotte et al. 2002; Bennetzen 2005; Morgante et al. 2007; Weil and Martienssen 2008; and see Slotkin et al. 2012, this volume). This diverse functional impact of TEs, and their intrinsic contribution to genomic plasticity, suggests that these elements play a major role in molecular diversification and, ultimately, species divergence.

In this chapter, we provide the reader with the fundamentals of TE biology, with an emphasis on plant elements. We begin with an overview of TE classification and transposition mechanisms, followed by an examination of the extensive variability in both inter- and intra-specific TE content across diverse plant taxa. Finally, we explore some of the general principles characterizing and influencing the genomic distribution, activity and evolution of TEs.

2.2 Transposable Element Classification

TEs can be broadly defined as DNA segments capable of chromosomal movement, either via replicative or conservative (cut-and-paste) mechanisms (discussed in more detail below). The TE classification system that we present here is similar to the one proposed by Wicker et al. (2007) and to the one implemented in Repbase, the most popular database of repetitive DNA sequences (http://www.girinst.org/). At the highest level, eukaryotic TEs comprise two major classes, and each class can be divided into subclasses based on their mechanism of chromosomal integration, which is reflective of the protein-coding capabilities and organizational structure of each class and subclass of elements (Figs. 2.1, 2.2). Class I elements, also known as retrotransposons, transpose via an RNA intermediate, which must be reverse transcribed prior to integration into the genome, while Class II elements transpose via a DNA intermediate (Finnegan 1989). Transposition of both classes of elements may result in a heritable increase in genomic copy number; hence, individual TE types are found in multiple copies (often referred to as a TE family) and comprise the majority of the repetitive fraction of eukaryotic genomes (e.g. Adams et al. 2000; The Arabidopsis Genome Initiative 2000; Lander et al. 2001; International Rice Genome Sequencing Project 2005). TEs have been found in virtually every organism studied to date (with few exceptions, such as *Plasmodium falciparum* and other Apicomplexa), although significant qualitative and quantitative variation abounds, even among closely related organisms (see below for a comparison among selected plant species).

The genomes of plants are packed with many and diverse TEs, and continue to serve as excellent models to yield some of the most significant advances in the field of transposon biology. The vast majority of repetitive DNA in the nuclear genomes of plants is derived from the proliferation of TEs, most often Class I RNA elements (Fig. 2.1) (e.g. SanMiguel et al. 1996; Vicient et al. 1999; Hawkins et al. 2006; Neumann et al. 2006; Vitte and Bennetzen 2006). Two major subclasses of Class I elements have been identified in plants: (1) Long terminal repeat (LTR) retrotransposons, whose reverse-transcription and subsequent integration as double-stranded DNA is mediated by an element-encoded reverse transcriptase and integrase, respectively, (2) non-LTR retrotransposons (sometimes called retroposons), which include long and short interspersed elements (LINEs and SINEs) and use target-primed reverse transcription, a mechanism coupling reverse transcription and integration. DIRS-like elements (named after *Dictyostelium* intermediate repeat sequence) represent a third subclass of retrotransposons integrated through an element-encoded tyrosine recombinase. They are relatively common in animals and fungi, but have yet to be found in flowering plants. Class II elements have been identified in every plant genome that has been thoroughly examined, and these can be divided in two major subclasses: (1) classic ‘cut-and-paste’ DNA transposons, characterized by terminal inverted repeats (TIRs), which are excised and reintegrated as double-stranded DNA by the action of an element-encoded transposase and (2) *Helitrons*, or rolling-circle transposons, which most likely transpose via a replicative mechanism involving a single-stranded DNA intermediate and which encode recombinase with Replicator initiator motif (Rep) and DNA Helicase domains (Fig. 2.1).

In plants, Class I elements (particularly LTR retrotransposons) make up the largest fraction of the TE complement (SanMiguel et al. 1996, 1998; Vicient et al. 1999;

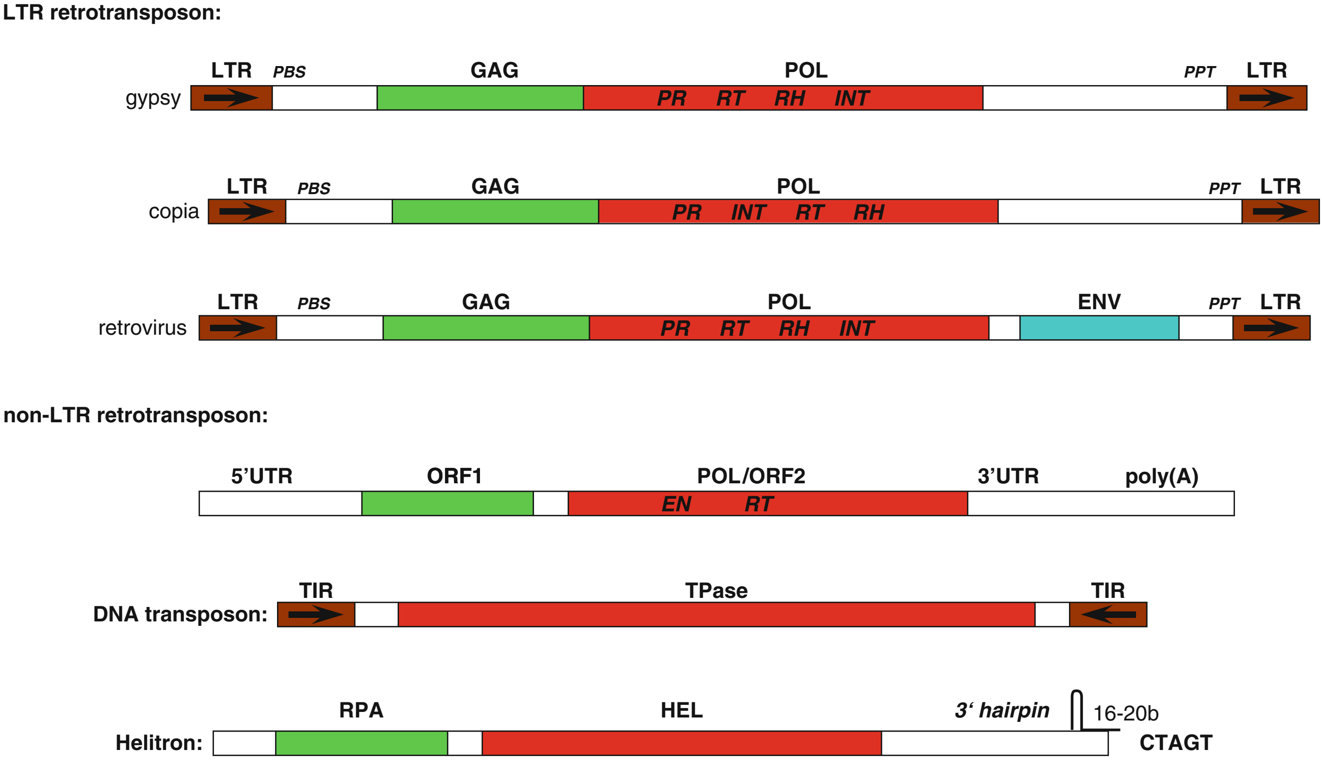

Fig. 2.1 Structure of main types of transposable elements. GAG and POL genes of LTR retrotransposons, ORF1 of non-LTR retrotransposons, transposase (TPase) of DNA transposons and replicative protein A (RPA) and helicase (HEL) of Helitrons are marked. Long terminal repeats (LTRs), primer-binding site (PBS) and polypurine tract (PPT) of LTR retrotransposons, 5′ UTR, 3′ UTR and poly(A) of non-LTR retrotransposons, terminal inverted repeats (TIR) of DNA transposons and 3′ hairpin of Helitrons are labeled. LTR retrotransposons are exemplified by *gypsy*, *copia* and retrovirus superfamilies. Protease (PR), reverse transcriptase (RT), RNaseH (RH), integrase (INT) and endonuclease (EN) domains are marked

Hawkins et al. 2006; Neumann et al. 2006; Vitte and Bennetzen 2006). The LTRs flanking a retrotransposon can range from just a few hundred base pairs to as much as 6 kb, and usually begin with 5′-TG-3′ and end with 5′-CA-3′. The LTR retrotransposons typically contain GAG and POL protein coding ORFs, which encode several enzymes (reverse transcriptase – RT; protease – PR; RNaseH – RH; integrase – INT) responsible for reverse transcription and integration of daughter sequences into new chromosomal locations. Two major superfamilies of LTR retrotransposons are found in plants, *gypsy*-like and *copia*-like (also known as *Metaviridae* and *Pseudoviridae*, respectively). Both types of LTR retrotransposons contain the same protein coding domains, but these are arranged in a different order. Their ancient origin is evidenced by the fact that they form deeply diverged monophyletic clades in phylogenetic analyses of reverse transcriptases (Eickbush and Malik 2002; Havecker et al. 2004). Non-LTR retrotransposons (LINEs and SINEs) are, as their name indicates, not flanked by LTRs, but complete LINEs can reach several thousand base pairs in length, contain coding sequences responsible for transposition, and often display a stretch of adenines or a simple sequence repeat at their 3′ end (Figs. 2.1, 2.2c).

Class II DNA elements are found in most eukaryotes, and despite their conservative transposition mechanism, have been capable of attaining relatively high copy numbers in some plants (see Sect. 2.3.1, Feschotte and Pritham 2007). Class II elements encode the machinery to facilitate their own transposition, usually in the form of a transposase (TPase) encoded by a single gene. "Cut-and-paste" transposition is associated with Subclass 1 DNA transposons, and occurs via TPase binding to the terminal inverted repeats (TIRs) of the element (Fig. 2.1), followed by excision and reintegration of the transposon at a new chromosomal location (Craig et al. 2002). The transposition mechanism of *Helitrons* has not been investigated in functional detail, but these elements are believed to employ a mechanism where only one DNA strand is cut, displaced and which serves as a template for replication of the element at a new locus (Kapitonov and Jurka 2007).

Both Class I and Class II TEs may be further divided into autonomous or non-autonomous elements dependent upon their ability to encode the enzymatic machinery responsible for movement. Non-autonomous elements may still be mobilized *in trans* if they retain the capacity to be recognized by the enzymes encoded by autonomous