The Pennsylvania State University

The Graduate School

Intercollege Graduate Degree Program in Plant Biology

PHYLOGENOMIC ANALYSIS OF ANCIENT GENOME DUPLICATIONS IN THE HISTORY OF PLANTS

A Dissertation in

Plant Biology

by

Yuannian Jiao

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2011

The dissertation of Yuannian Jiao was reviewed and approved* by the following:

Claude dePamphilis
Professor of Biology
Dissertation Advisor
Chair of Committee

Hong Ma
Professor of Biology

John Carlson
Professor of Molecular Genetics

Webb Miller
Professor of Comparative Genomics and Bioinformatics

Naomi Altman
Professor of Statistics

Teh-hui Kao
Chair of Plant Biology Graduate Program

*Signatures are on file in the Graduate School

ABSTRACT

Whole-genome duplication (WGD), or polyploidy, followed by gene loss and diploidization, has generally been viewed as a primary source of material for the origin of evolutionary novelties. Most flowering plants have been shown to be ancient polyploids that have undergone one or more rounds of WGDs early in their evolution, and many lineages have since undergone additional, independent and more recent genome duplications. It was suggested that the paleopolyploidy events were crucial to the radiation and success of angiosperms, but evidence for proposed ancient genome duplications remains equivocal. Plant genomes are highly dynamic and usually go through intense structural rearrangements and gene loss following duplication. Old(er) WGDs can not be detected using within-genome colinearity method. Although the occurrence of WGD event(s) is well accepted, the actual number, phylogenetic timing, and age of the event(s) remain equivocal. Here we mainly employed phylogenomic approach to address following questions: 1) To more precisely time and evaluate the fates of genes following genome duplications in eudicots, we built a phylogenomic pipeline to reconstruct the evolutionary relationships of 4433 gene families from the complete gene sets of *Arabidopsis, Populus, Vitis,* and *Oryza*. For each family, we determined the gene duplications relative to the evolutionary tree of the organisms, and also matched gene pairs of specific WGD events for later synonymous substitution (Ks) analysis. Concentrations of duplication events at shared branch points in the gene family phylogenies confirmed some of the results of earlier studies, and also indicated new interesting signals for polyploidy events. 2) To track more ancient WGDs pre-dating the divergence of monocots and eudicots, we used comprehensive phylogenomic analyses of nine sequenced plant genomes and more than 12.6 million new expressed-sequence-tag (EST) sequences from phylogenetically pivotal lineages. Two additional, previously unnoticed ancient WGDs have been elucidated in the common ancestor of extant seed plants dated at ~ 319 million

years ago (mya) and extant angiosperms dated at ~ 192 mya, respectively. Significantly, these ancestral WGDs resulted in the diversification of regulatory genes important to seed and flower development, suggesting that they were involved in major innovations that ultimately contributed to the rise and eventual dominance of seed plants and angiosperms. 3) To precisely time the γ event, which was proposed before the separation of all rosids, we mapped the duplication events onto phylogenetic trees that include the paralogs created by the γ event and orthologous genes from other species. The overwhelming majority of well-resolved *Vitis* duplications were placed before the separation of rosids and asterids and after the split of monocots and eudicots, providing evidence for the WGD (γ) early in eudicot evolution. A large proportion of the *Vitis* duplications were placed after the divergence of basal eudicots, supporting the γ triplication was likely restricted to core-eudicots. Global gene family phylogenies are a valuable complement to genome-scale structural analysis, incorporating extensive genome-wide evidence even without a sequenced genome, and facilitate a better understanding of WGD events in plants.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank all people who has helped me and encouraged me during my doctoral study.

I owe my deepest gratitude to my advisor Dr. Claude dePamphilis for his great support and guidance for the past five years. His patience, enthusiasm, and immense knowledge had motivated me all the time of research and writing of this thesis. I could not have come to this far without his inspiration and great efforts. Besides my advisor, I would like to thank my thesis committee members, Dr. Hong Ma, Dr. Webb Miller, Dr. Naomi Altman, Dr. John Carlson for their insightful comments and suggestions. I also would like to thank all my coauthors, Jim Leebens-Mack, Douglas Soltis, Pamela Soltis, Haiying Liang, Sandra Clifton, Scott Schlarbaum, Stephan Schuster and all members in they lab. I would like to thank Norman Wickett, Joshua Der, Eric Wafula, Loren honaas, Yan Zhang, Yuchen Zhang, Yeting Zhang, Jill Duarte, Lena Landherr, Paula Ralph, Kerr Wall from dePamphilis' lab, and Xiaofan Zhou and other members in Ma's lab for their numerous valuable suggestions and generous help.

I want to thank my wife, Zi Shi, for her support, consideration and encouragement not only in my daily life, but also academically. I thank my friends in the USA: Junlei Sun, Yufan Zhang, Chenwei Ma, Yi Liu, Hui Zhang, Guang Sheng, Xiaoying Meng, Xuan Ma, Yili Sun and Fang Cong, for all the joy we have had in the past years.

Finally, and most importantly, I would like to thank my parents for their boundless love and tremendous assistance throughout my life. To them I dedicate this thesis.

Chapter 1 Introduction

Brief history of gene duplication research

Genomic complexity is driven, to a large extent, by gene duplication, retention, and divergence (Ohno, 1970; Lynch and Conery, 2003). More than 40 years ago, Susumu Ohno proposed that gene duplication was the single most important factor in evolution (Ohno, 1967). Three years later (Ohno, 1970), he reiterated that without duplicated genes, many of the major evolutionary transitions of interest to humans would have been impossible. He argued that the complex organisms could not have evolved just by the modification of existing genes, without the creation of new ones by duplication(Ohno, 1970). It has been generally accepted that gene duplications have played necessary roles in providing the raw material for the evolution of the species on earth. The earliest studies of gene duplication focused on observable changes in organismal phenotypes and chromosome morphology (Tischler, 1915). In the late 1960s, the development of starch gel electrophoresis enabled scientists to use isozyme electrophoresis studies to identify gene duplicates in polyploids (Stuber and Goodman, 1983). However, most of the surveys were restricted to a set of approximately 30 enzyme loci that could be readily stained and visualized in starch or acrylamide gels. In 1993 Kary Mullis was awarded the Nobel Prize in Chemistry for inventing of the polymerase chain reaction (PCR), which fueled revolutionary advances in molecular biology. With PCR, pairs of gene duplicates could be rapidly identified, for example, three related zebrafish Noggin genes were discovered using a single set of primers (Furthauer et al., 1999). Advances in sequencing technology (the Maxam-Gilbert method and later the Sanger method) enabled relatively fast and cost efficient DNA sequencing, providing better understanding of gene duplication and divergence. However, sequencing an entire genome

was still difficult and time consuming, as well as prohibitively expensive for most eukaryotic genomes. Over the last decade, "next-generation sequencing" (NGS) technologies have been developed, such as Roche/454, Illumina/Solexa, SOLiD. These revolutionary technologies deliver fast, inexpensive and accurate genome information. Therefore, the whole genome sequences of many organisms are now becoming available in large public databases including the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov), Ensembl (www.ensembl.org), J. Craig Venter Institute (www.jcvi.org), Joint Genome Institute (www.jgi.doe.gov), and phytozome (www.phytozome.net). These sequences make it possible to study gene duplication or genome duplication using an exclusively bioinformatic-based approach now. More and more genome sequences for multiple branches of virtually all angiosperm clades, including major crops and/or botanical models should be expected during the next decade. These sequences will provide essential information for relating gene or even genome-level duplications to aspects of morphological and physical variation that have contributed to the widespread success of various plant groups (eg. Angiosperms).

Mechanisms of gene duplication

In general, four mechanistic modes of gene duplication have been proposed. Although all modes can duplicate a gene, they play different roles in the dosage relationship between new duplicated gene and gene dosages for the rest of the genome. The four modes are as follows: 1) polyploidy (whole genome duplication or WGD), 2) chromosomal segmental duplication, 3) tandem duplication, and 4) transposition. Here we do not include gene gains by horizontal gene transfer (HGT) to be gene duplication even if they may result in the expansion of a gene family. The mechanism by which a pair of duplicated genes is generated can be inferred by their genomic context (Vision et al., 2000; Simillion et al., 2002; Blanc et al., 2003; Bowers et al., 2003; Cannon et al., 2004). The number of types of polyploidy and the characteristic of each type of

polyploidy has long been debated. In general, two major types of polyploidy have historically been recognized: autopolyploidy and allopolyploidy (Kihara, 1926). Autopolyploids are polyploids with multiple chromosome sets derived from a single species. Autopolyploidy can involve hybridization between populations of the same species, which can arise from a spontaneous, naturally occurring genome doubling, or following fusion of 2n gametes (unreduced gametes). Allopolyploids are polyploids with chromosomes derived from different species, which is the result of doubling of chromosome number in an F1 hybrid. Allopolyploidy has long been assumed much more common than autopolyploidy (Spring, 2003). Genes duplicated by whole genome duplication tend to remain for a while in syntenic blocks with other duplicated genes of the same age, as approximated by the rates of synonymous substitutions per site (Ks).

Segmental duplications are those in which many genes and their upstream regions are duplicated in a single event, which are proposed from the spontaneous duplication of large DNA segments, both intra- and inter-chromosomal (Koszul et al., 2004). Duplicated genes that reside next to one another are referred to as tandem duplicates. All genomes have tandem repeats. It has been suggested that unequal crossover between homologous chromosomes during meiosis and possible unequal sister-chromatin exchange during mitosis might have played important roles in generating tandem duplicates (Brown and Dawid, 1968; Cronn et al., 1996). Inversions within tandem arrays are common, so after some evolutionary time, tandem duplicates are not necessarily adjacent genes. Finally, a single gene transposition duplication means genes have transposed from ancestral site to another chromosomal position, such as transposon mediated transposition. Transpositions lead to duplications because even if a transposed gene leaves its site of origin in transposition process, the segregants include duplication gametes.

The evolutionary significance of whole genome duplications

Among the four different modes of gene duplication, polyploidy has been recognized as an especially powerful driver of gene creation. A change in ploidy is typically expected to be deleterious and an evolutionary dead end (Otto and Whitton, 2000). However, many organisms have descended from ancestors who doubled their genomes either through autopolyploidy or allopolyploidy. WGD is especially common in flowering plant lineages, with all angiosperms having at least one detectable genome duplication event in their evolutionary history (Blanc and Wolfe, 2004; Jaillon et al., 2004; Cui et al., 2006; Tang et al., 2008; Tang et al., 2009; Jiao et al., 2011). Although polyploidy is much less frequent in animals than in plants, hundreds of known insects and vertebrate species are likely polyploid, including amphibians and fish (Christoffels et al., 2004; Jaillon et al., 2004; Kuraku et al., 2009). It has been observed that WGDs have sometimes given rise to species rich groups, such as >25,000 species of fish and >350,000 species of flowering plants (Van de Peer et al., 2009; Jiao et al., 2011). Genome duplication could increase species diversity through reciprocal gene loss, subfunctionalization and speciation (Lynch and Force, 2000; Soltis, 2009). It has been reported that ~1700 (8%) ancestral loci of *Tetraodon nigroviridis* and zebrafish underwent reciprocal gene loss, which would be sufficient to result in reproductive isolation and contributed to speciation events that occurred after the teleost WGD (Semon and Wolfe, 2007).

In addition, it has been argued that genome duplication could reduce the risk of extinction through functional redundancy, mutational robustness, and increased rates of evolution and adaption (Crow and Wagner, 2006). The most compelling evidence is that a majority of the independent genome duplications in many different plant lineages, including legumes, cereals, Solanaceae, lettuce and cotton, are clustered in time ~60-70 million years ago (mya). This wave of WGDs coincides with the Cretaceous-Tertiary (KT) extinction event, suggesting that polyploidy plants may have outcompeted their diploid progenitors in the face of dramatically

changing environmental conditions (Fawcett et al., 2009). Several other studies also indicated that polyploidy plants have a higher tolerance of a wide range of environmental conditions in comparison to their diploid relatives (Thompson and Lumaret, 1992; Ramsey, 2011). In stable ecosystems, diploid ancestors might show better adaptation than newly formed polyploids. However, whenever the ecosystems are severely perturbed, polyploids (after WGD) might have had a pronounced selective advantage over their diploid sister species for the survival and long-term evolutionary successes (Van de Peer et al., 2009).

Fate of the duplicated genes

It has been proposed that most duplicated genes will be silenced and eventually be eliminated due to degenerative mutations in coding and/or regulatory elements; this process is named "nonfunctionalization" (Lynch and Conery, 2000). Although nonfunctionalization could not provide raw genetic material for divergence, reciprocal gene loss (the most common consequence of duplication) after WGD in separated populations might genetically isolate these populations as discussed above (Lynch and Force, 2000). However, reciprocal gene loss –prime suspects for agents of plant speciation- has not yet been well documented on a whole-genome scale in plants, in contrast to yeasts (Scannell et al., 2006; Scannell et al., 2007).

In addition, there are many important exceptions. Those genes which are not lost are referred to as retained. The retained duplicates are of interest. The DDC (duplication-degeneration-complementation) model was proposed to explain a common fate of duplicates, called "subfunctionalization". Degenerative mutations in regulatory elements controlling the expression of two duplicated genes lead to complementary expression patterns (Force et al., 1999) that if combined together, would comprise the ancestral expression pattern. Later on, a protein coding subfunctionalization model was proposed, in which duplicates are generated by

alternative splicing (Yu et al., 2003) or asymmetric mutations that decrease the efficacy of each gene and lead to selection for genomes to retain both (Edelmann et al., 2001). The DDC model (and related variants) is currently the most popular model to explain duplicate gene retention.

Another possibility is that duplicated genes can evolve new functions - "neofunctionalization" (Ohno, 1970; Lynch and Conery, 2000). The two copies of the duplicated gene are originally redundant, which means they perform the same function and that inactivation or mutation of one gene should have no effect on the biological phenotype. This is the classic positive selection model to explain retention (Ohno, 1970). This mechanism is thought to be relatively infrequent if considering the rapid fractionation process that follows genome duplication (Langham et al., 2004). However, the evolution of novelty involves this process eventually. Subfuncitonalization followed by functional divergence can lead to new functions – eg, sub-neofunctionalization. In addition, it has been shown that exonization of intron sequences and pseudoexonization of exon sequences, just by one point deletion or insertion, have contributed to the divergence of duplicated MADs-box genes in sequence structure and gene function (Xu and Kong, 2007).

A study of regulatory genes in *Arabidopsis* suggested a fourth fate for duplicated genes – "*hypofunctionalization"* – a case in which one member of a duplicate pair is consistently expressed 2-3 fold lower than the other member of the duplicates, potentially as protection against loss of function through redundancy (Duarte et al., 2006). This kind of alteration of gene expression pattern also has been observed in synthetic allotetraploids (Wang et al., 2006). By comparing global gene expression profiles in a synthetic *Arabidopsis* allotetraploid, formed by combing *A. arenosa* with *A. thaliana*, strong expression dominance was observed for the *A. arenosa* parent, coupled with suppression of the *A.thaliana* genome (Wang et al., 2006).

It is important to consider why some duplicated genes might be preserved in the genome over long evolutionary time periods while the others do not. Duplicates from different modes of duplication (tandem, polyploidy, or segmental) tend to be retained in a biased (non-random) pattern. It has been found that the most recent tetraploidy in the *Arabidopsis* lineage preferentially retained genes encoding transcription factors, protein kinases, transferases (Maere et al., 2005; Freeling, 2009; Jiao et al., 2011). This is also consistent with patterns of gene retention following WGD in vertebrates (Kassahn et al., 2009).

Several other models have been proposed to explain the biased retention pattern following duplication besides the subfunctionalization and neofunctionalization mentioned above. Conserved gene retention idea is from observations in *Saccharomyces cerevisiae* and *Caenorhabditis elegans* (Davis and Petrov, 2004). Conserved genes tend to be retained after duplication but newer genes are not. Another idea relates gene retention probability to gene expression level. In yeast, it has been observed that selection for increased levels of gene expression was a significant factor determining which genes were retained and which were returned to single-copy state (Seoighe and Wolfe, 1999). Finally, Freeling *et al.* (Freeling and Thomas, 2006; Freeling, 2009) proposed a balanced gene drive model, which derives from the gene balance hypothesis (Papp et al., 2003; Veitia, 2004; Birchler and Veitia, 2007). None of the models mentioned above but this one could predict or explain that more "connected" genes (eg. by protein-protein interaction) are more likely to be retained after WGD than after tandem duplications. Otherwise, it will change the gene dosage, which will confer a loss of fitness (haploinsufficiency). A balanced gene drive model could also explain why gene families encoding protein kinases, motors, and transcription factors tend to be retained after WGD. It has also been argued that balanced gene drive should tend to drive up morphological complexity,

which is potentiated by duplicated functional models (gene networks). However, this hypothesis could be difficult to test (Semon and Wolfe, 2007).

How whole genome duplication is studied

Identification of block duplications

Several methodologies have been proposed and widely used to unravel genome duplication. Identification of duplicated blocks of genes in genome sequences provides the strongest evidence of ancient polyploidy (Vision et al., 2000; Bowers et al., 2003; Jaillon et al., 2007; Tang et al., 2008). This method depends on a within-genome comparison that aims to delineate regions of conserved gene content and order in different parts of the genome, and has been widely used in plants (Bowers et al., 2003; Jaillon et al., 2007; Tang et al., 2008) and vertebrates (Goodstadt and Ponting, 2006; Scannell et al., 2007). Gene order in vertebrates is very conserved after hundreds of millions of years of divergence (Dehal and Boore, 2005). However, much more rapid structural evolution have been shown in flowering plants, estimated to have diverged 125-235 mya (Bowers et al., 2003; Jaillon et al., 2007; Lyons et al., 2008; Tang et al., 2008). BLASTZ-Chain-Net, Lagan-Supermap, PipMaker and VISTA pipelines are commonly used in vertebrate genome alignments, focusing mainly on identifying orthologous regions without much concern about confusion with paralogous regions (Mayor et al., 2000; Schwartz et al., 2000; Kent, 2002; Frazer et al., 2003). In general, “all vs all” BLASTP results are used as inputs to build a homology matrix. Syntenic blocks are uncovered by clustering neighboring matches inside the matrix. ADHoRe (Vandepoele et al., 2002) and DiagHunter (Cannon et al., 2003) are based on a similar approach to detect the signal of WGDs in plants. Two recently developed packages, DAGchainer (Haas et al., 2004) and ColinearScan (Wang et al., 2006),

formulate the problem by dynamic programming. When candidate collinear regions have been detected, usually some sort of empirical or statistical test is performed to calculate the probability that the observed colinearity could have been generated by chance, which effectively improve sensitivity and specificity of colinearity prediction. These approaches to detect WGDs have been "bottom-up", in which one start with the most recent duplication, then merge the defined blocks as ancestral pseudochromosomes to further find evidence for second recent duplication and so on. However, it could be difficult to apply to angiosperms (flowering plants) because of frequent genome duplications and subsequent extensive genomic rearrangements and gene loss. In 2008, an alternative and complementary "top-down" algorithm (MCscan) was developed to combine related pairwise collinear segments into one inferred order based upon multiple colinearity (Tang et al., 2008). For example, A-B and C-D were identified as two pairwise syntenic blocks. If B and C share the same chromosomal region, the final output would be A-(BC)-D. This is common in plants. By using MCscan, it has been demonstrated that triplicated structures are over most of the genome of papaya and grapevine, suggesting a triplication event shared at least by all rosids (Jaillon et al., 2007; Ming et al., 2008). Finally, although this syntenic approach provided compelling evidence for genome-level duplications, it needs to have completed genome sequences and has been proven to be difficult to track very ancient genome duplications in plants due to extensive genome rearrangements and gene loss. Further challenges will be discussed in later chapters.

Age distributions of duplicated genes

Another approach to infer WGD is to build age distributions of paralogs, where the number of paralogs is plotted against their age. Because synonymous substitutions are generally exposed to little selection, synonymous divergence (Ks) can be considered a proxy for time, thus

allowing the age of a duplication event to be estimated. A genome wide duplication event simultaneously generates thousands of paralogous pairs with initial Ks values of zero; as time passes, gene pairs diverge and exceptional peaks in the distribution of Ks values interpreted as large-scale or genome duplications (Lynch and Conery, 2000; Blanc et al., 2003; Blanc and Wolfe, 2004; Cui et al., 2006). This method does not depend on genomic positional information, and can be used in any species for which moderately large gene sequence sets are available including ESTs. However, depending on the level of gene loss, real paralogous pairs may not be easily identified based only on sequence similarity to one another, particularly if the species in question has undergone more than one round of genome duplications. Gene death and divergence processes can also obscure the detection of *bona fide* genome duplications using the Ks approach (Cui et al., 2006), and peaks in gene duplication are not necessarily due to genome duplication (see for example, Fig 2K in (Blanc and Wolfe, 2004)). In addition, the rate of synonymous substitutions (Ks) between two homologous sequences will become saturated when the two sequences have diverged sufficiently. Therefore, it is difficult for Ks methods to detect genome duplications that are so old that saturation of synonymous substitutions would occur.

In addition, it has been demonstrated that molecular clocks run at different speeds in different angiosperms. For example, a hexaploidy event was at least shared by all rosids, including *Arabidopsis*, *Populus*, *Carica*, and *Vitis* (Tang et al., 2008). The median Ks between *Vitis* gamma paleologs is 1.22, which is much lower than that of *Carica* (1.76) and *Populus* (1.54). *Arabidopsis* has a still-faster rate of evolution. The Ks distribution of *Arabidopsis* could only show two peaks which corresponding the alpha (0.86) and beta (1.8) respectively. The median Ks of gamma duplicates is close to saturation (Ks ca. 2.0), and is much larger than those of gamma duplicates in the other three rosid species (Tang et al., 2008). Therefore, the nucleotide substitution rates of these four eudicots are significantly different, with the *Arabidopsis* lineage

evolving fastest. This analysis indicates that no universally applicable molecular clock exists, at least within eudicots. When using this approach to study WGD, further supporting information is needed to time the age of putative WGD(s).

Duplication histories from phylogenomic analysis of gene families

Another way to identify genome duplication is by phylogenetic analysis of gene families (Bowers et al., 2003; Blomme et al., 2006). Gene family histories reflect both speciation events and gene duplication events; when the species relationships are known, gene duplications can often be inferred. In this way, genome-wide gene family histories will reflect the phylogenetic timing of surviving gene duplications in the organisms whose genomes are compared. A genome-wide duplication event should result in a large excess of gene duplications in comparison to those produced by individual duplication events. By interpreting all of the trees, the relative position of enriched duplications in the species tree can be used to infer the existence and timing of a large-scale or even whole-genome wide duplication event. This approach has some advantages compared with Ks and synteny analyses: extensive gene loss and genome shuffling (rearrangements, chromosomal fusions and fissions) are the two biggest challenges for Ks and synteny methods, but less so in a phylogenetic approach. A Phylogenomic approach can in principle incorporate evidence from all surviving genes within gene families, while synteny analyses requires syntenic blocks of large size. However, a limitation of phylogenomic analysis is uncertainty in the phylogenetic reconstructions; the number of sequenced plant genomes is still relatively small, and gene tree phylogenies may be subject to artifacts such as long-branch attraction. For this reason, care is required to use outgroups that are not too distant, to carefully select taxon sampling to cut potential long branches, and to rely upon methods of phylogenetic reconstruction (eg. maximum likelihood) that can be consistent in the face of heterogeneous rates

of sequence evolution (Bergsten, 2005). Phylogenetic timing of duplication events without additional supporting evidence is also inconclusive; a concentration of duplications at a specific point in time may only identify a time period with an accelerated duplication of individual genes or multi-gene segments rather than indicating WGD. Therefore, the most conclusive evidence should take into account multiple sources of evidence from gene families, phylogenetic trees, Ks analysis, patterns of gene retention, and genomic synteny.

Current research on dating duplication events in flowering plants

Ancient polyploidy is common in plant lineages. Complete or near-complete genome sequences are now available for several flowering plants: *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Carica papaya*, *Populus trichocarpa*, *Cucumis sativus*, *Glycine max*, *Theobroma cacao,* Strawberry (*Fragaria vesca*)*,* apple *(Malus x domestica Borkh.), Vitis vinifera*, and the graminoid monocots *Oryza sativa*, *Brachypodium distachyon*, *Zea mays*, and *Sorghum bicolor.* Primary descriptions are in progress for at least three more that we are aware of (*Medicago truncatula*, *Mimulus guttatus*, *Aquilegia coerulea*), and sequencing is in progress for many more (eg. *Amborella trichopoda*). The next decade will see essentially completed sequences for all major angiosperm clades, including major crops and/or botanical models, and multiple gymnosperm lineages. Two plant genomes from seedless land plants (*Selaginella moellendorffii* and *Physcomitrella patens*) are also available, and can serve as outgroups for comparative analyses involving flowering plants. Phylogenetic relationship among these plants is shown in Figure 1-1.

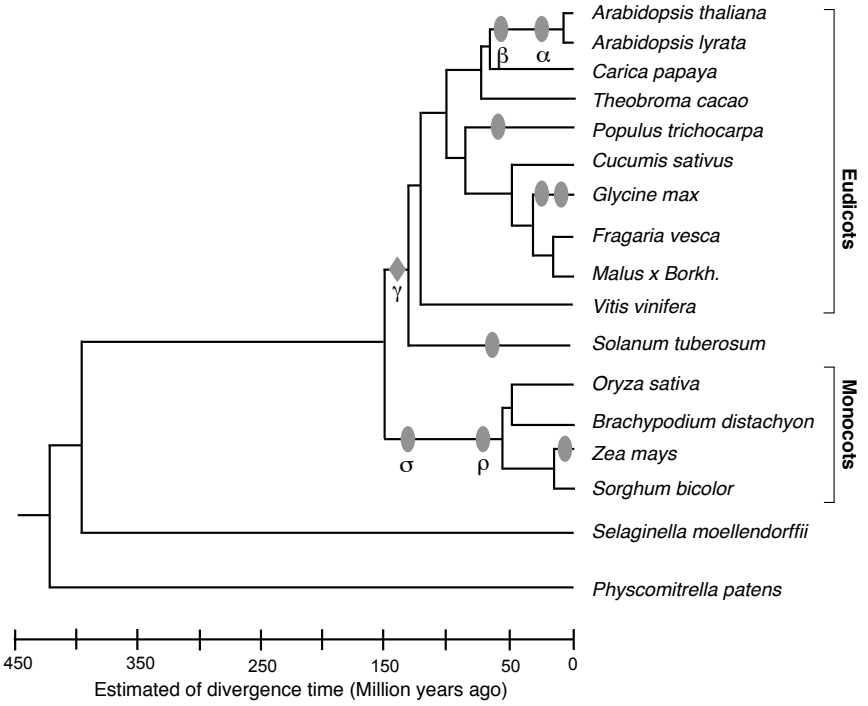


Figure 1-1. Phylogenetic tree of plants with a sequenced genome.
WGDs, inferred from syntenic analyses of sequenced genomes, are indicated by ovals. The hexaploidy event was indicated by diamond.

By exploring syntenic blocks in *Arabidopsis* and then determining the relationship of the gene pairs by a simplified phylogenetic approach, Bowers *et al.*'s study (Bowers et al., 2003) suggested two WGDs in the history of *Arabidopsis* lineage after the divergence of monocots and dicots: an older one (β') that was shared by all eudicots, and estimated to have occurred 170-235 Mya (million years ago), and a younger one (α') is Brassica-wide, and estimated to have occurred 14.5-20.4 Mya ago (hereafter in this chapter, α', β', and γ' refer to WGD events proposed by Bowers *et al.* 2003, whereas α, β, and γ refer to WGD events proposed by Tang *et al.* 2008). The *Populus trichocarpa* (poplar) genome paper estimated the accumulated nucleotide divergence for multiple genes from syntenic blocks at fourfold synonymous third-codon transversion position (4DTV) values and proposed a WGD event likely shared by rosids besides the most recent Salicoid one in *Populus* lineage (Tuskan et al., 2006). However, recent studies on the genome of

Vitis vinifera (winegrape) and *Carica papaya* suggested a different scenario (Jaillon et al., 2007; Ming et al., 2008; Tang et al., 2008; Tang et al., 2008). These studies concluded that the previously identified WGD shared by rosids (or eudicots) was a triplication event (named γ), because most *Vitis* syntenic block regions have two other paralogous regions. This triplicated genome structure is also evident in a recently published genome sequence of *Theobroma cacao* (Argout et al., 2011). They also proposed that there are two recent genome duplication events affecting *Arabidopsis* within the crucifer lineage, suggested by one *Vitis* region corresponding to four *Arabidopsis* segments. These two recent genome duplication events are not shared by *Carica papaya*, as one *Vitis* region only tends to match one *Carica* block. Another controversy concerns the exact phylogenetic position of *Vitis*, either as a basal rosid (Jansen et al., 2006; Velasco et al., 2007) or rosid I close to *Populus* (Jaillon et al., 2007). When the genome sequence of soybean (*Glycine max*) was completed, two additional genome duplications after γ were identified, a soybean-lineage-specific palaeotetraploidization (at approximately 13 mya), and an early-legume duplication (59 mya) shared with *Medicago* (Schmutz et al., 2010). A recent WGD (>50 mya) has also been demonstrated in the apple genome after the divergence of populus and apple (Velasco et al., 2010). No recent WGD was found in the Cucumber or Carica genome, which could explain the relatively small number of genes within both genomes (Cucumber: 26,682 genes; Carica: 24,746 genes). The diploid strawberry genome also lacks recent large-scale duplication. In addition, the signature of the ancient γ has also been eroded by chromosome rearrangement and genome size reduction in the strawberry genome (Shulaev et al., 2011).

In monocots, two WGD events have been identified after the divergence of monocots and eudicots, and before the origin of the grasses, as reflected in the four sequenced grass genomes (Paterson et al., 2004; Yu et al., 2005; Wu et al., 2008; Tang et al., 2009). *Sorghum* reflects no grass-specific WGDs (Paterson et al., 2009), but its' close relative maize is a paleotetraploid

(Gaut and Doebley, 1997). By analysis of Ks distribution for paralogous genes of *Acorus americanus*, three components were identified in the paralogous pairs by the mixture model approach, which might represent two distinct large-scale duplication events (Cui et al., 2006). Cui *et al.* also identified three mixture components in the Ks distribution of the basal angiosperm *Nuphar advena,* which provided evidence of ancient polyploidy events. However, only one component was identified for a total of 69 *Amborella* paralogous pairs, consistent with no ancient polyploidy in the history of *Amborella* using this limited gene set (Cui et al., 2006; Soltis, 2009).

Contents of this dissertation

In this dissertation, we employed phylogenomic approach to track and time ancient WGDs in the history of plants by using sequences from completed plant genomes and species with moderately large ESTs. The second chapter is about WGDs in Eudicots. The two recent duplication events (α and β) are evident in the history of *Arabidopsis* lineages after the separation with *Populus*. Phylogenomic evidence identified that the ancient γ event was also shared with asterids, which was not determined by synteny analyses due to lack of completed asterid genome. Concentrations of gene duplications also suggested potential WGD events in the lineages leading to Solanaceae and to Asteraceae, but not across all Asteridae. The third chapter mainly focused on more ancient WGDs pre-dating the divergence of monocots and eudicots. Two additional, previously unnoticed ancient WGDs have been elucidated in the common ancestor of extant seed plants and in the common ancestor of extant angiosperms. These two ancient events were potentially involved in major innovations that ultimately contributed to the rise and eventual dominance of seed plants and angiosperms. The fourth chapter is mainly focused on precisely timing the hexaploidy event (γ). Completed genome sequences and large sets of unigenes obtained from transcriptomes of various Asteridae, basal eudicots, non-grass monocots,

magnoliids and basal angiosperms were used to estimate the phylogeny of gene families in which *Vitis* gene sets located on syntenic blocks in the *Vitis* genome were grouped. We provided evidence for the γ event early in eudicot evolution, most likely restricted to core-eudicots. Global gene family phylogenies are a valuable complement to genome-scale structural analysis, incorporating extensive genome-wide evidence even without a sequenced genome, and facilitate a better understanding of WGD events in plants.

References

Argout X, Salse J, Aury JM, 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 JS, 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

Bergsten J (2005) A review of long-branch attraction. Cladistics **21:** 163-193

Birchler JA, Veitia RA (2007) The gene balance hypothesis: from classical genetics to modern genomics. Plant Cell **19:** 395-402

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res **13:** 137-144

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667-1678

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol **7:** R43

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438

Brown DD, Dawid IB (1968) Specific gene amplification in oocytes. Oocyte nuclei contain extrachromosomal replicas of the genes for ribosomal RNA. Science **160:** 272-280

Cannon SB, Kozik A, Chan B, Michelmore R, Young ND (2003) DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. Genome Biol **4:** -

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. BMC Plant Biol **4:** 10

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol **21:** 1146-1151

Cronn RC, Zhao X, Paterson AH, Wendel JF (1996) Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. J Mol Evol **42:** 685-705

Crow KD, Wagner GP (2006) What is the role of genome duplication in the evolution of complexity and diversity? Mol Biol Evol **23:** 887-892

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res **16:** 738-749

Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol **2:** E55

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3:** e314

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Mol Biol Evol **23:** 469-478

Edelmann L, Stankiewicz P, Spiteri E, Pandita RK, Shaffer L, Lupski JR, Morrow BE (2001) Two functional copies of the DGCR6 gene are present on human chromosome 22q11 due to a duplication of an ancestral locus. Genome Res **11:** 208-217

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci USA **106:** 5737-5742

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531-1545

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: A review of methods and available resources. Genome Res **13:** 1-12

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu Rev Plant Biol **60:** 433-453

Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res **16:** 805-814

Furthauer M, Thisse B, Thisse C (1999) Three different noggin genes antagonize the activity of bone morphogenetic proteins in the zebrafish embryo. Dev Biol **214:** 181-196

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci USA **94:** 6809-6814

Goodstadt L, Ponting CP (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol **2:** e133

Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics **20:** 3643-3646

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome

duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature **431:** 946-957

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

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol **6:** 32

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97-100

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009) Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res **19:** 1404-1418

Kent WJ (2002) BLAT - The BLAST-like alignment tool. Genome Res **12:** 656-664

Kihara O (1926) Chromosomenzahlen und systematische Gruppierung der Rumex-Arten Cell and Tissue Research **4:** 475-481

Koszul R, Caburet S, Dujon B, Fischer G (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. Embo J **23:** 234-243

Kuraku S, Meyer A, Kuratani S (2009) Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? Mol Biol Evol **26:** 47-59

Langham RJ, Walsh J, Dunn M, Ko C, Goff SA, Freeling M (2004) Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166:** 935-945

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155

Lynch M, Conery JS (2003) The origins of genome complexity. Science **302:** 1401-1404

Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459-473

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M (2008) Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol **148:** 1772-1781

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci USA **102:** 5454-5459

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I (2000) VISTA: visualizing global DNA sequence alignments of arbitrary length. Bioinformatics **16:** 1046-1047

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, 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 ML, 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 JK, 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 MC, 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

Ohno S (1967) Sex Chromosomes and Sex-linked Genes. Heidelberg, Springer-Verlag

Ohno S (1970) Evolution by gene duplication. Springer-Verlag

Otto SP, Whitton J (2000) Polyploid incidence and evolution. Annu Rev Genet **34:** 401-437

Papp B, Pal C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature **424:** 194-197

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 KF, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:** 551-556

Paterson AH, Bowers JE, Chapman BA (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proc Natl Acad Sci USA **101:** 9903-9908

Ramsey J (2011) Polyploidy and ecological adaptation in wild yarrow. Proc Natl Acad Sci USA **108:** 7096-7101

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature **440:** 341-345

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci USA **104:** 8397-8402

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 XC, 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

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker - A web server for aligning two genomic DNA sequences. Genome Res **10:** 577-586

Semon M, Wolfe KH (2007) Consequences of genome duplication. Current Opinion in Genetics & Development **17:** 505-512

Semon M, Wolfe KH (2007) Reciprocal gene loss between *Tetraodon* and *zebrafish* after whole genome duplication in their ancestor. Trends in Genetics **23:** 108-112

Seoighe C, Wolfe KH (1999) Yeast genome evolution in the post-genome era. Current Opinion in Microbiology **2:** 548-554

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 JM, Rees DJ, Williams KP, Holt SH, Ruiz Rojas JJ, Chatterjee M, Liu B, Silva H, Meisel L, Adato A, Filichkin SA, Troggio M, Viola R, Ashman TL, Wang H, Dharmawardhana P, Elser J, Raja R, Priest HD, Bryant DW, Jr., Fox SE, Givan SA, Wilhelm LJ, Naithani S, Christoffels A, Salama DY, Carter J, Lopez Girona E, 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

Simillion C, Vandepoele K, Van Montagu MC, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci USA **99:** 13627-13632

Soltis DE, Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A., Zheng, C., Sankoff, D, dePamphilis, C.W., Wall, P.K. and Soltis, P.S. (2009) Polyploidy and angiosperm diversification. American Journal of Botany**:** 13

Spring J (2003) Major transitions in evolution by genome fusions: from prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates. J Struct Funct Genomics **3:** 19-25

Stuber CW, Goodman MM (1983) Inheritance, intracellular localization, and genetic variation of phosphoglucomutase isozymes in maize (*Zea mays L.*). Biochem Genet **21:** 667-689

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

Tang H, Bowers JE, Wang X, Paterson AH (2009) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci USA **107:** 472-477

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

Thompson JD, Lumaret R (1992) The evolutionary dynamics of polyploid plants: origins, establishment and persistence. Trends Ecol Evol **7:** 302-307

Tischler G (1915) Chromosomenzahl, form und individualität in planzenreiche. Progr Rei Bot **5:** 164

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 GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin 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, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, 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, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, 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

Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet **10:** 725-732

Vandepoele K, Saeys Y, Simillion C, Raes J, Van de Peer Y (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res **12:** 1792-1801

Veitia RA (2004) Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. Genetics **168:** 569-574

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 CE, 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 x domestica Borkh.*). Nat Genet **42:** 833-839

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, Dematte 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

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

Wang JL, Tian L, Lee HS, Wei NE, Jiang HM, Watson B, Madlung A, Osborn TC, Doerge RW, Comai L, Chen ZJ (2006) Genomewide nonadditive gene regulation in *Arabidopsis* allotetraploids. Genetics **172:** 507-517

Wang X, Shi X, Li Z, Zhu Q, Kong L, Tang W, Ge S, Luo J (2006) Statistical inference of chromosomal homology based on gene colinearity and applications to *Arabidopsis* and rice. BMC Bioinformatics **7:** 447

Wu Y, Zhu Z, Ma L, Chen M (2008) The preferential retention of starch synthesis genes reveals the impact of whole-genome duplication on grass evolution. Mol Biol Evol **25:** 1003-1006

Xu GX, Kong HZ (2007) Duplication and divergence of floral MADS-box genes in grasses: Evidence for the generation and modification of novel regulators. Journal of Integrative Plant Biology **49:** 927-939

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li S, Li X, Zheng H, Cong L, Lin L, Yin J, Geng J, Li G, Shi J, Liu J,

Lv H, Li J, Wang J, Deng Y, Ran L, Shi X, Wang X, Wu Q, Li C, Ren X, Wang J, Wang X, Li D, Liu D, Zhang X, Ji Z, Zhao W, Sun Y, Zhang Z, Bao J, Han Y, Dong L, Ji J, Chen P, Wu S, Liu J, Xiao Y, Bu D, Tan J, Yang L, Ye C, Zhang J, Xu J, Zhou Y, Yu Y, Zhang B, Zhuang S, Wei H, Liu B, Lei M, Yu H, Li Y, Xu H, Wei S, He X, Fang L, Zhang Z, Zhang Y, Huang X, Su Z, Tong W, Li J, Tong Z, Li S, Ye J, Wang L, Fang L, Lei T, Chen C, Chen H, Xu Z, Li H, Huang H, Zhang F, Xu H, Li N, Zhao C, Li S, Dong L, Huang Y, Li L, Xi Y, Qi Q, Li W, Zhang B, Hu W, Zhang Y, Tian X, Jiao Y, Liang X, Jin J, Gao L, Zheng W, Hao B, Liu S, Wang W, Yuan L, Cao M, McDermott J, Samudrala R, Wang J, Wong GK, Yang H (2005) The Genomes of *Oryza sativa*: a history of duplications. PLoS Biol **3:** e38

Yu WP, Brenner S, Venkatesh B (2003) Duplication, degeneration and subfunctionalization of the nested *synapsin-Timp* genes in *Fugu*. Trends Genet **19:** 180-183

Chapter 2 Tracking the history of genome duplications in flowering plants: evidence from global gene family phylogenies

There is strong evidence that the ancestors of major eudicot lineages have undergone one or more rounds of whole-genome duplication (WGD) following the divergence of monocots and eudicots. Although the occurrence of WGD event(s) is well accepted, the actual number, phylogenetic timing, and age of the event(s) remain equivocal. To address these issues, we built a phylogenomic pipeline to reconstruct the evolutionary relationships of 4433 gene families from the complete gene sets of *Arabidopsis*, *Populus*, *Vitis*, and *Oryza*. 1787 families were characterized by a surviving duplication shared by rosid I (*Populus*) and rosid II (*Arabidopsis*). These alignments were populated with unigenes of Asteridae and re-estimated the phylogenies to track potential WGD event(s) in eudicots, rosids, and asterids. Very little evidence was found to support large-scale duplications shared only by rosid I and rosid II, rejecting prior hypotheses of a rosid-wide WGD. The overwhelming majority of resolved duplications shared by rosid I/II were placed before the separation of rosids and asterids, providing evidence for WGD (γ') early in eudicot evolution. Concentrations of gene duplications also suggested potential WGD events in the lineages leading to Solanaceae and to Asteraceae, but not across all Asteridae. Finally, our results support two rounds of WGD (α' and β') in the *Arabidopsis* lineage after the divergence of rosid I/II. Global gene family phylogenies are a valuable complement to genome-scale structural analysis, incorporating extensive evidence even without conservation of gene order or a sequenced genome, and facilitate a better understanding of WGD events in eudicots.

Background

Gene duplication provides raw genetic material for the evolution of functional novelty and is considered to be a driving force in evolution (Ohno, 1970; Adams and Wendel, 2005). Four mechanisms of gene duplication have been proposed: whole-genome duplication (WGD), segmental duplication, tandem duplication, and transposition (Zhang, 2003). The mechanism by which a pair of duplicated genes is generated can often be inferred from their genomic context (Vision et al., 2000; Blanc et al., 2003; Bowers et al., 2003; Cannon et al., 2004). WGDs, which involve the doubling of the entire genome, are of special importance and have been well documented in several lineages including fungi (Kellis et al., 2004), flowering plants (Vision et al., 2000; Blanc et al., 2003; Bowers et al., 2003; Tang et al., 2008; Tang et al., 2008; Van de Peer et al., 2009), and vertebrate animals (Christoffels et al., 2004; Jaillon et al., 2004; Dehal and Boore, 2005). Studies in these lineages support an association between WGD and resulting gene duplications (Blanc et al., 2003; Cui et al., 2006), functional divergence in duplicate gene pairs (Duarte et al., 2006; Johnson and Thomas, 2007), phenotypic novelty (Conrad and Antonarakis, 2007), and potentially rapid increases in species diversity (De Bodt et al., 2005; Meyer and Van de Peer, 2005).

Several methodologies have been proposed and widely used to unravel genome duplication. Identification of syntenic blocks of genes in genome sequences provides strong evidence of segmental duplication or even ancient polyploidy (Bowers et al., 2003; Jaillon et al., 2007; Tang et al., 2008). This method depends on genomic positional information, and has been used in some plants (Bowers et al., 2003; Cannon et al., 2003; Jaillon et al., 2007; Lyons et al., 2008; Tang et al., 2008) and vertebrates (Goodstadt and Ponting, 2006; Scannell et al., 2007). Recently, several comparative genomic approaches were introduced to aid the identification and comparison of homologous regions from multiple genomes, such as MCSCAN (Tang et al., 2008; Tang et al., 2008) and CoGe (Lyons et al., 2008). However, challenges are expected when

applying syntenic block analysis to angiosperms because of likely frequent genome duplication events and subsequent extensive genomic rearrangements and gene loss. Another approach is to evaluate the frequency distribution of per-site synonymous divergence (K_s) for pairs of duplicate genes, where *synonymous site* divergence is a proxy for the *age* of the duplication event, as synonymous substitutions are largely immune to strong selective pressures (Lynch and Conery, 2000; Blanc et al., 2003; Blanc and Wolfe, 2004; Maere et al., 2005; Cui et al., 2006). A genome-wide duplication event simultaneously generates thousands of paralogous pairs; those remaining pairs tend to correspond to peaks in the distribution of K_s values. Therefore, such K_s peaks strongly suggest past genome or other large-scale duplications. This method does not depend on genomic positional information, and can be used in any species for which moderately large EST sets are available (Blanc and Wolfe, 2004; Cui et al., 2006). However, depending on the level of gene loss, paralogous pairs might not be easily identified using sequence similarity, particularly if the species in question has undergone more than one round of genome duplication or if the genes are incompletely sampled. Gene death and divergence can obscure the detection of *bona fide* genome duplications using the K_s approach (Cui et al., 2006), and peaks in the K_s distribution are not necessarily due to whole-genome duplication (see for example, Figure 2K in (Blanc and Wolfe, 2004), where a peak in the K_s distribution is due to an ancient amplification of tandemly repeated genes). In addition, the rate of synonymous substitutions (K_s) between two homologous sequences becomes saturated when the two sequences have diverged sufficiently. For this reason, some analyses have focused on K_a ("dN") divergences, given that non-synonymous substitution will remain unsaturated to much high genetic divergence levels (Vision et al., 2000). Conversely, gene pairs may have little or no divergence in very recent genome duplications, making it difficult for K_s methods to detect very young or very old genome duplications.

Another approach to identify genome duplication events is through the use of genome-scale phylogenetic analysis of gene families (Bowers et al., 2003; Blomme et al., 2006; Jansen et al., 2006). Although individual genes may be lost in some phylogenies, a broader picture can be drawn from simultaneous consideration of many or all gene families. By interpreting these trees, the relative position of the duplicated genes on the tree can be used to determine the timing of a duplication event. A phylogenomic approach can, in principle, incorporate evidence from all surviving genes within gene families without requiring large syntenic blocks, whereas synteny analyses requires this positional information. However, a limitation of phylogenomic analysis is uncertainty in the phylogenetic reconstructions; the number of sequenced and annotated plant genomes is still relatively small, and gene tree phylogenies may be subject to artifacts such as long-branch attraction (Felsenstein, 1978). For these reasons, care is required to select outgroups that are not too distant, to base analyses on multiple taxa, and to rely upon methods of phylogenetic reconstruction (e.g., maximum likelihood) that can be consistent in the face of heterogeneous rates of sequence evolution (Felsenstein, 1978). Phylogenetic timing of duplication events may also be inconclusive; a concentration of duplicate genes at a specific point in time could reflect a period of accelerated duplication of individual genes or multigene segments rather than indicating WGD. Therefore, a thorough understanding of the timing and mechanism of ancient genome-scale duplication should include evidence from gene family phylogenies, K_s analysis, and genomic synteny. Sampedro *et al.* demonstrated how evidence could be integrated across all three approaches to study the evolutionary history of the EXPANSIN gene family (Sampedro et al., 2005). Here, we seek to evaluate how genome-wide gene family phylogenies can contribute to our understanding of WGD in cases where structural evidence has been inconclusive or sequenced genomes are unavailable.

Angiosperms (flowering plants) are by far the largest group of land plants, consisting of more than 300,000 species (Crane et al., 1995); most species are included in the two large groups,

monocots and eudicots. Eudicots are, in turn, composed of two large groups, rosids and asterids, which contain most of the commonly known eudicot species and all of the sequenced angiosperms (*Arabidopsis*, *Populus*, *Vitis*, *Carica*) other than the sequenced grass family monocots (*Oryza*, *Zea*, and *Sorghum*). The rosid and asterid lineages are further subdivided into two clades each, rosid I/II, and asterid I/II. By exploring syntenic blocks in *Arabidopsis* (rosid II) and then determining the relationship of the gene pairs by a simplified phylogenetic approach, Bowers *et al.* (Bowers et al., 2003) suggested two WGDs after the divergence of monocots and eudicots: one (β) appeared to be eudicot-wide, estimated to have occurred 170-235 Myr (million years) ago, and the other (α) was much more recent, estimated to have occurred 14.5-20.4 Myr ago within the family Brassicaceae. In *Populus trichocarpa* (poplar, rosid I), the accumulated nucleotide divergence for duplicated gene blocks was estimated using fourfold synonymous third-codon transversion positions (4DTV); these estimates of sequence divergence should be relatively transparent to selection, and suggested that there might have been a WGD event shared by rosids in addition to a more recent "salicoid" WGD in the *Populus* lineage (Tuskan et al., 2006). This rosid-wide WGD would have also been shared by *Arabidopsis* (rosid II); however, it was not detected by Bowers *et al.*(Bowers et al., 2003), and could be interpreted as the β event with an uncertain age estimate, or as an independent rosid-wide event.

Recently, analyses of the genomes of *Vitis vinifera* (winegrape, grapevine) and *Carica papaya* (papaya tree) suggested a different scenario for the nature of the early eudicot WGD event (Jaillon et al., 2007; Lyons et al., 2008; Tang et al., 2008; Tang et al., 2008). These studies suggested that the previously identified whole-genome duplication shared by rosids (or eudicots) was a triplication event, as most syntenic blocks of *Vitis* have two other paralogous regions. They also proposed two recent genome duplication events (α' and β') (hereafter α, β and γ refer to WGD events as proposed by Bowers *et al.* (Bowers et al., 2003), whereas α', β' and γ' refer to WGD events proposed by Tang *et al.* (Tang et al., 2008; Tang et al., 2008); see list of

abbreviations) in the history of *Arabidopsis* within the crucifer lineage, as suggested by the detection of four *Arabidopsis* segments that correspond to only one *Vitis* region. However, previous studies have indicated only one round of WGD (α event) in the history of the *Arabidopsis* lineage after divergence of rosid I and rosid II (Vision et al., 2000; Bowers et al., 2003).

Genome sequences of several rosids have since been completed, and EST resources within the asterids have also grown extensively (Childs et al., 2007). Very large EST datasets from multiple members of Asteraceae (e.g., *Helianthus annuus,* sunflower) and Solanaceae (e.g., *Solanum tuberosum*, potato), in particular, provide good coverage of the gene sets from the two largest asterid lineages. These provide substantial resources for using the strategy of relative dating (Bowers et al., 2003). In this strategy, a large-scale phylogenetic analysis of gene families is used to place duplication events before or after the divergence of certain lineages; gene pairs that map to the phylogeny at specific times can be classified for later analysis of synteny and K_s. These phylogenomic and relative dating approaches are used here to more precisely time the shared rosid I and rosid II WGD event, as well as unravel genome duplication events within asterids.

The aims of our study were to investigate: (1) whether the shared rosidI/rosidII polyploidy event (γ') was also shared with asterids, and is therefore at least as old as the core eudicots (= asterids + rosids), (2) whether there is evidence supporting large-scale duplications in asterids, and (3) whether there is more than one round of WGD (α' and β') in the history of *Arabidopsis* after the divergence of rosid lineages I and II. To answer these questions, we constructed gene families by using complete genome sequences of four representative species (*Arabidopsis*, *Populus*, *Vitis* and *Oryza*) as well as EST sequences of several asterid species, and then bioinformatically interpreted their resulting phylogenies (Figure 2-1). Based on our results, the shared rosid I and rosid II polyploidy event (γ') was also shared with asterids. In asterids, we

found very few asterid-wide duplications, but a large number of duplications in Solanaceae and Asteraceae, respectively. In addition, our phylogenomic evidence appears to support two rounds of WGD (α' and β') in the lineage leading to *Arabidopsis* after the divergence from the common ancestor shared with the *Populus* lineage, supporting the Tang *et al.* hypothesis using this independent approach.

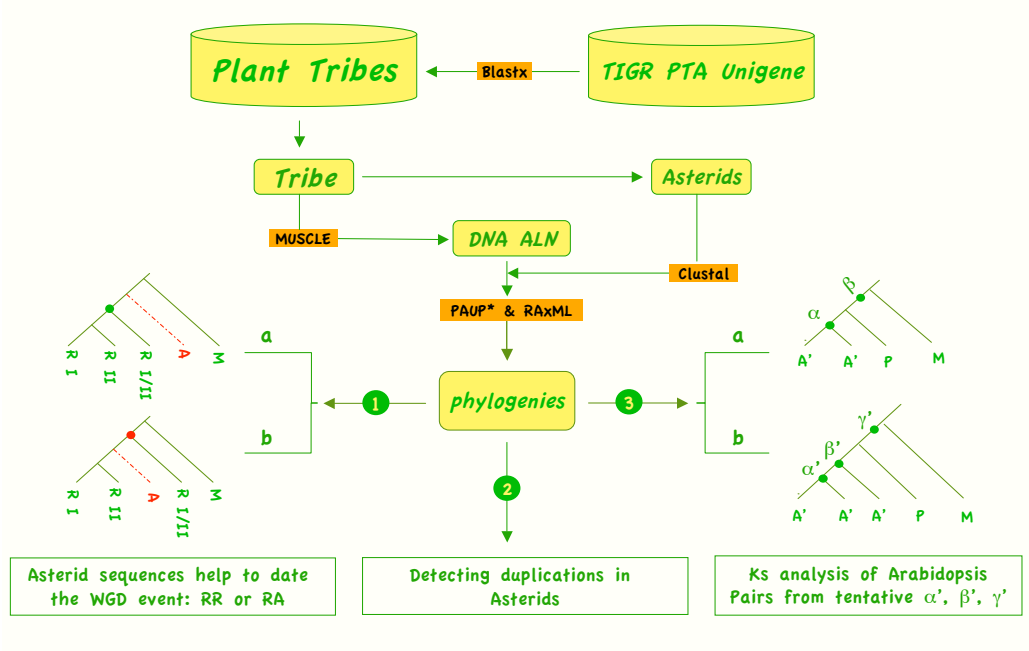


Figure 2-1. Schematic diagram detailing main flows of data analysis.

Three routes address three issues: route #1 = Exemplar tree topologies illustrating the potential timing of shared gene duplications in rosid-wide RR (a) or eudicot-wide RA (b); route #2 = Detecting duplications in asterids (asterid-wide (AA), asterid I (A1A1), and asterid II (A2A2)); route #3 = Two different topologies supporting only one major duplication event for the *Arabidopsis* lineage after the split of rosid I and rosid II (a) or two WGDs (b). Note: A' = *Arabidopsis* (rosid II); P = *Populus* (rosid I); M = monocots; R = rosids; A = asterid; A1 = asterid I; A2 = asterid II.

Results and Discussion

Initial screen with four genomes

In order to identify gene families with a phylogenetic signal that could potentially aid in timing ancient genome duplications among major eudicot lineages, we downloaded putative gene families, called tribes, from the PlantTribes database (http://fgp.huck.psu.edu/tribedb/) (Wall et al., 2008). A total of 4433 tribes containing at least one *Oryza* (monocots, as outgroup), one *Arabidopsis* (rosid II) and one *Populus* (rosid I) sequence were identified. The total number of genes included in these tribes is presented in Table 2-1. Phylogenetic trees were constructed for each tribe, using the Maximum Likelihood (ML) optimality criterion as implemented in RAxML (Stamatakis et al., 2005), and Maximum Parsimony (MP) and Neighbor Joining (NJ) methods as implemented in PAUP*4.0b10 (Wilgenbusch and Swofford, 2003). Gene trees for the sequenced genomes were generally well-resolved and topologies were largely congruent, regardless of the phylogenetic method with which the tree was reconstructed.

Table 2-1. Summary of genes and analyzed genes for four sequenced plant genomes included in this study.

Analyzed gene number is the number of genes contained in the selected tribes having at least one *Arabidopsis* gene, one *Populus* gene, and one *Oryza* gene.

Species	Annotation version	Annotated genes	Analyzed genes
Arabidopsis (*Arabidopsis thaliana*)	TAIR version 7	26784	14533
Poplar (*Populus trichocarpa*)	JGI version 1.1	45554	22687
Grape (*Vitis vinifera*)	Genoscope release	30434	14699
Rice (*Oryza sativa*)	RAP release 2	29389	15213

For this initial screen of 4433 gene trees, we defined an orthogroup as a subclade of the gene family (tribe) phylogeny that includes a monocot outgroup sequence and all sister eudicot sequences (including eudicot duplications, if any). A large number of orthogroups were observed with shared duplications of *Arabidopsis* and *Populus* genes (hereafter referred to as an AP DUP event). Figure 2-2A, for example, is a common ML topology of a tribe (Tribe 2786) where two major clades have survived from the shared rosid I + rosid II duplication (node indicated by red

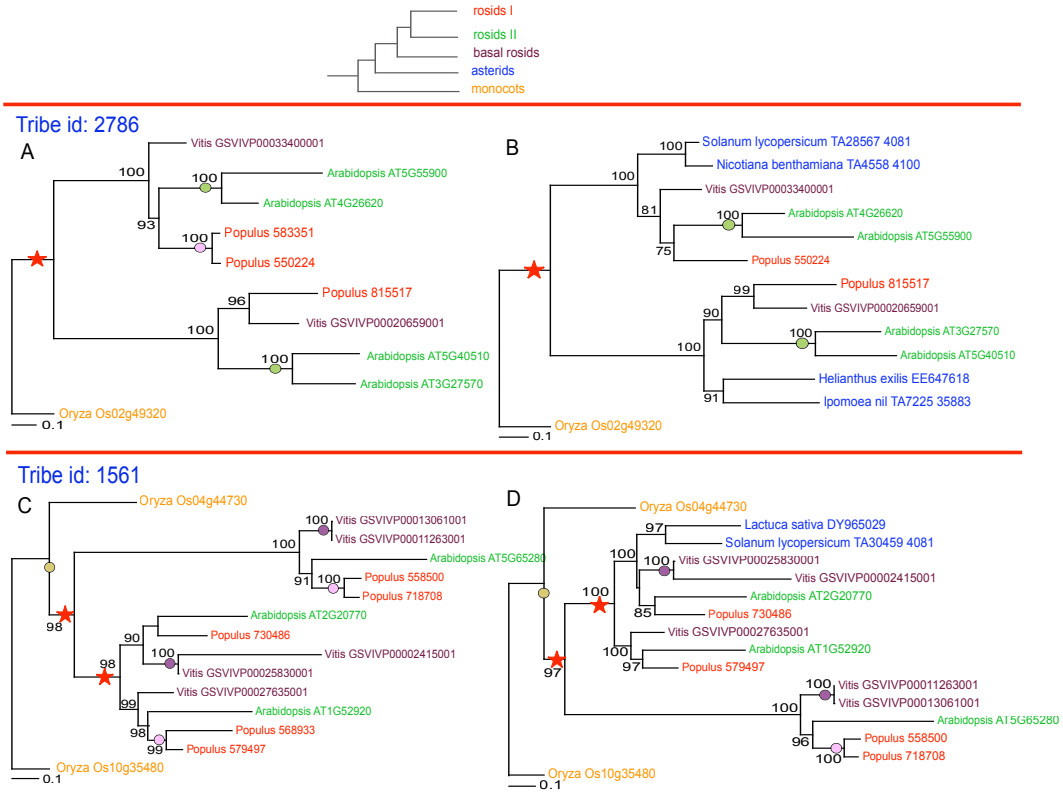


Figure 2-2. Exemplar ML phylogenies

(A) Common ML topology of a tribe (Tribe 2786) where two major clades have survived the shared rosid I and rosid II duplication. (B) Common ML phylogeny of the tribe (Tribe 2786) with asterid sequences added whose duplication pattern is consistent with Fig 1 route #1b. (C) Exemplar ML phylogeny of a tribe (Tribe 1561) whose topology is potentially consistent with triplication. (D) Common ML phylogeny of the tribe (Tribe 1561) with asterid sequences whose topology is consistent with eudicot-wide polyploidy. Legend: Star = eudicot-wide duplication; colored circles = recent independent duplications; numbers = bootstrap support values.

star), as well as additional duplicates from recent duplications (indicated by colored circles). Figure 2-2C is another exemplar ML phylogeny of a tribe (Tribe 1561), where three major clades have survived from the shared rosid I and rosid II duplications or triplication event. The total number of orthogroups identified differed slightly by phylogenetic reconstruction methods (1617 [shared *Arabidopsis*-*Populus* duplications] / 5020 [orthogroups identified] in ML, 1589/5235 in MP, 1866/5134 in NJ with bootstrap support greater than 50) (Table 2-2). These duplicated orthogroups were included in 2010 gene trees (Tribes). After the split of rosid I and rosid II, we found 1879 (ML), 1689 (MP), and 1699 (NJ) clades in the *Arabidopsis* lineage and 3435 (ML), 4829 (MP), and 4569 (NJ) clades in the *Populus* lineage, with good bootstrap support (> 80) for recent gene duplications (Figure 2-3, node 1'+1 and 2). The large concentrations of duplicated genes suggest that polyploidy events most likely occurred on the branch subtending the specified nodes on these trees. Some error in the sequence annotation or even in the gene trees is to be expected in a genome-scale analysis, but the overall signal from so many independent analyses of gene families should not be positively misleading. For example, somewhat truncated sequences will generally decrease bootstrap support for the placement of such sequences, or even cause an unresolved topology, resulting in an undercount of well-supported duplicated clades.

Table 2-2. Summary of orthogroups showing AP (*Arabidopsis* + *Populus*) duplications inferred from four genome gene trees using three phylogenetic methods.

BS>=80 and BS>=50 are counts of nodes resolved with bootstrap values >=80 or >=50, respectively. Total = the number of orthogroups studied.

ORTHO	ML		MP		NJ	
	BS>=80	BS>=50	BS>=80	BS>=50	BS>=80	BS>=50
TOTAL	4279	5020	4191	5235	4142	5134
AP DUP	1181	1617	723	1589	970	1866
PERCENT	27.6%	32.2%	17.3%	30.4%	23.4%	36.3%

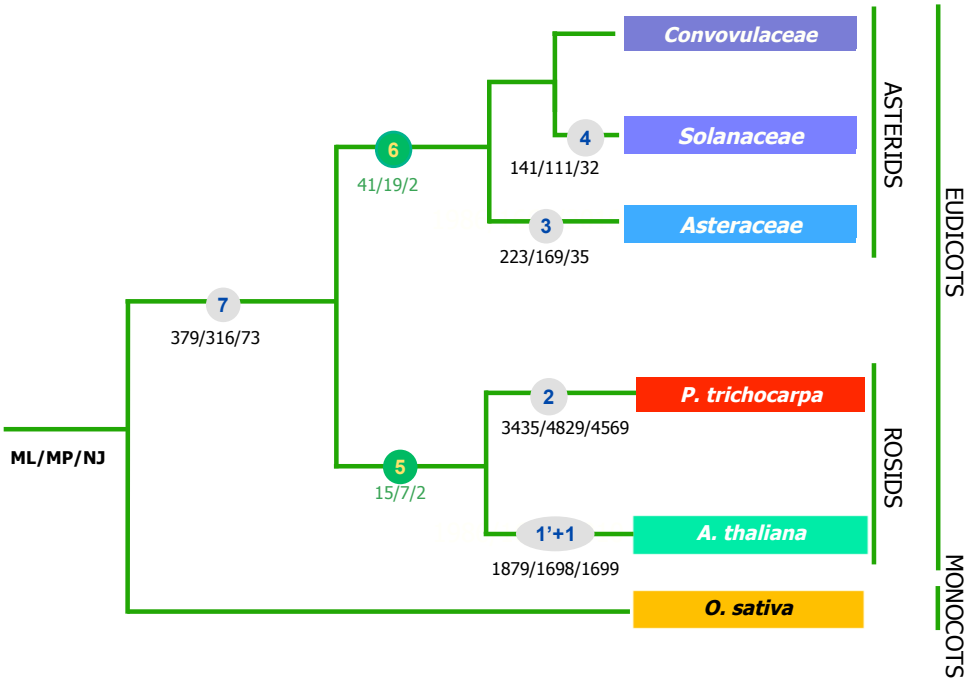


Figure 2-3. Phylogentic timing of inferred gene duplications.

Values given are: number of orthogroups showing duplications at the specified node on trees generated with Maximum Likelihood/Maximum Parsimony/Neighbor Joining with bootstrap value greater than 80. Gray circles indicate nodes for which there is strong support for WGDs; green circles indicate nodes for which there is very little evidence of WGDs.

We excluded *Vitis* when screening for orthogroups with a shared rosid I and rosid II duplication because of the uncertainty of *Vitis'* placement: either rosid I (Jaillon et al., 2007) or basal rosid (Jansen et al., 2006; Velasco et al., 2007). According to our phylogenomic results, a nearly equal number of clades support each of the two positions (data not shown). Without further testing the phylogenetic position of *Vitis*, which may require broader taxon sampling, we excluded *Vitis* data in the process of counting duplicated orthogroups, which has the effect of decreasing the number of duplicated orthogroups (without counting ((APV)V) and ((AV)V) and ((PV)V)), without producing a false-positive signal.

Rosid-wide or Eudicot-wide large-scale duplication events?

The 2010 tribes with at least one shared *Populus* (rosid I) and *Arabidopsis* (rosid II) duplication (AP DUP) identified a likely polyploidy event at least as old as the common ancestor of rosid I and rosid II. To establish whether this large-scale duplication was restricted to rosids (Figure 2-1 route #1a, RR DUP) or shared with asterids (i.e. eudicot-wide) (Figure 2-1 route #1b, RA DUP), these tribes were then populated with unigenes of multiple Asteridae species (Table 2-3). Asterid sequences were identified for 1787 of the 2010 tribes and were subsequently used for phylogenetic analysis. For example, tribe 2786 was populated with four asterid unigenes, two of which were clustered into each of the surviving rosid I and rosid II clades, supporting RA DUP rather than RR DUP (Figure 2-2B). Alternatively, only two asterid unigenes were clustered into tribe 1561, and grouped to one of the three duplicated clades, also supporting RA DUP (Figure 2-2D). Regardless of the phylogenetic method used, the number of topologies that identify a shared rosid and asterid duplication (RA) greatly exceeds the number of topologies that identify rosid-wide duplications (RR) (Table 2-4, Figure 2-3). A eudicot-wide (RA) duplication is supported (bootstrap >80) by 379 orthogroups using ML, 316 using MP, and 73 using NJ, whereas only 15, 7, and 2 orthogroups, respectively, identify a rosid-wide (RR) duplication. If stringency is lowered to bootstrap support > 50, there are 623 orthogroups using ML, 530 using MP, and 248 using NJ that show an RA duplication, and 46, 28, 10 that show an RR duplication (Table 2-4). Therefore, there is strong support in our phylogenomic analysis for a WGD shared by asterids and rosids.

Table 2-3. Summary of unigene sequences of Asteridae included in this study.

Legend: # EST = total number of ESTs in TIGR PTA database; # Unigenes = total number of unigenes in TIGR PTA database; # Involved = total number of unigenes assembled in the tribes with one or more rosid I and rosid II duplications.

FAMILY	GENUS SPECIES (COMMON NAME)	Tax ID	# EST	# Unigenes	# Involved
Convolvulaceae (Morning-glory)	*Ipomoea nil* (Morning glory)	35883	61354	22340	2885
Convolvulaceae (Morning-glory)	*Ipomoea batatas* (Sweetpotato)	4120	17805	9167	1197
Convolvulaceae (Morning-glory)	*Ipomoea trifida*	35884	1379	908	108
Solanaceae (Nightshade)	*Capsicum annuum* (Pepper)	4072	30830	15404	2029
Solanaceae (Nightshade)	*Nicotiana benthamiana*	4100	26955	10162	1398
Solanaceae (Nightshade)	*Nicotiana langsdorffii x Nicotiana sanderae*	164110	12443	7202	804
Solanaceae (Nightshade)	*Nicotiana sylvestris* (S. American tobacco)	4096	7818	6775	1104
Solanaceae (Nightshade)	*Nicotiana tabacum* (Cultivated tobacco)	4097	72915	38612	3464
Solanaceae (Nightshade)	*Petunia x hybrida* (Garden petunia)	4102	10197	6613	838
Solanaceae (Nightshade)	*Solanum chacoense* (Chaco potato)	4108	6570	5932	923
Solanaceae (Nightshade)	*Solanum habrochaites*	62890	7997	4305	605
Solanaceae (Nightshade)	*Solanum lycopersicum* (Tomato)	4081	200248	45585	6180
Solanaceae (Nightshade)	*Solanum pennellii*	28526	8336	3772	557
Solanaceae (Nightshade)	*Solanum tuberosum* (Potato)	4113	219485	81072	10590
Asteraceae (Daisy)	*Cichorium intybus* (Chicory)	13427	3356	2457	427
Asteraceae (Daisy)	*Gerbera hybrid cultivar*	18101	16859	9482	1140
Asteraceae (Daisy)	*Helianthus annuus* (Common sunflower)	4232	93279	44662	5810
Asteraceae (Daisy)	*Helianthus argophyllus* (Silverleaf sunflower)	73275	34552	21194	2923
Asteraceae (Daisy)	*Helianthus exilis* (Serpentine sunflower)	400408	32818	22759	3469
Asteraceae (Daisy)	*Helianthus paradoxus* (Paradox sunflower)	73304	10310	6421	1189
Asteraceae (Daisy)	*Helianthus petiolaris* (Prairie sunflower)	4234	27456	15712	2165
Asteraceae (Daisy)	*Lactuca perennis* (Perennial lettuce)	43195	29108	13716	2066
Asteraceae (Daisy)	*Lactuca sativa* (Garden lettuce)	4236	80681	33115	6649
Asteraceae (Daisy)	*Lactuca serriola* (Prickly lettuce)	75943	55450	24949	3678
Asteraceae (Daisy)	*Lactuca virosa* (Wild lettuce)	75947	30033	13617	1997
Asteraceae (Daisy)	*Senecio aethnensis*	121540	1880	1436	232
Asteraceae (Daisy)	*Senecio cambrensis* (Welsh ragwort)	285720	2108	1667	267
Asteraceae (Daisy)	*Senecio chrysanthemifolius*	121541	2014	1540	271
Asteraceae (Daisy)	*Senecio squalidus* (Oxford ragwort)	121554	1921	1472	237
Asteraceae (Daisy)	*Senecio vulgaris* (Ragwort)	76276	1951	1578	245
Asteraceae (Daisy)	*Stevia rebaudiana* (Stevia)	55670	5385	3993	716
Asteraceae (Daisy)	*Taraxacum kok-saghyz* (Rubber dandelion)	333970	4702	3479	570
Asteraceae (Daisy)	*Taraxacum officinale* (Dandelion)	50225	41258	19701	2860
Asteraceae (Daisy)	*Zinnia elegans* (Zinnia)	34245	17914	15952	1654
Total number of Asteridae sequences			**1177367**	**516751**	**71247**

Table 2-4. Summary of orthogroups showing different types of duplication inferred from gene trees using three phylogenetic methods.

BS>=80 and BS>=50 are counts of nodes resolved with bootstrap values >=80 or >=50, respectively. Total = the number of orthogroups studied. RA DUP and RR DUP refer to node 7 and node 5 of Figure 2-3 respectively with duplications at the specified branch point.

ORTHO	ML		MP		NJ	
	BS>=80	BS>=50	BS>=80	BS>=50	BS>=80	BS>=50
TOTAL	1239	1519	913	1243	459	674
RA DUP	379	623	316	530	73	248
PERCENT	30.6%	41.0%	34.6%	42.6%	15.9%	36.8%
RR DUP	15	46	7	28	2	10
PERCENT	1.2%	3.0%	0.8%	2.3%	0.4%	1.5%

In addition to our phylogenomic analyses, we also examined published gene trees. We limited our consideration to only those gene families that were well populated with asterid genes as well as those from sequenced rosid genomes. The four sub-classes of the MADS box gene family, *APETALA1*, *APETALA3*, *AGAMOUS*, and *SEPALLATA*, all suggest a eudicot-wide duplication (Litt and Irish, 2003; Kramer et al., 2004; Zahn et al., 2005). The Class III HD-Zip family has 5-10 members, encodes homeodomain-leucine zipper transcription factors and regulates meristem initiation, vascular and leaf development; the gene tree for this gene family also indicates a eudicot-wide duplication (Prigge and Clark, 2006). The RPB2 family encodes the second largest subunit of RNA polymerase II. Both ML and MP RPB2 gene trees indicate the occurrence of an RPB2 gene duplication early in the evolution of eudicots (Luo et al., 2007). Table 2-5 summarizes the results of other independent, published phylogenies of gene families; the duplication signal strongly favors shared core eudicot duplications. For the published trees, there is no signal detected for ancient shared duplications restricted to all rosids or restricted to all asterids.

Table 2-5. Summary of published gene families showing duplication patterns relevant to this study.

R: rosids; A: asterids. No evidence is seen for a rosid-wide duplication pattern, while each gene family provides support for eudicot-wide [or earlier] duplications.

Gene family	Duplication Pattern	References
APETALA1 (A class)	((RA)((RA)(RA)))	Litt, A. *et al*, 2003
APETALA3 (B class)	((RA)(RA))	Kim, S. *et al*, 2004
AGAMOUS (C/D class)	((RA)(RA))	Kramer, E. M. *et al*, 2004
SEPALLATA (E class)	((RA)((RA)(RA)))	Zahn, L. M. *et al*, 2005
Glutamine synthetase (GS)	((RA)((RA)(RA))) & ((RA)(RA))	Biesiadka *et al*, 1997
Class III HD-Zip	((RA)(RA))	Prigge, M.J. *et al*, 2006
RNA polymerase II	((RA)(RA))	Luo, J. *et al*. 2007

This shared core eudicot polyploidy has been proposed as a triplication event by syntenic block analysis (Jaillon et al., 2007; Tang et al., 2008; Tang et al., 2008). Employing one monocot taxon as an outgroup, we identified 287 orthogroups from ML trees with more than two paralogous clades of rosid I and II sequences, such as the (((AP)(AP))(AP)) topology (note: A=*Arabidopsis*, P=*Populus*), which might indicate the hypothesized triplication event (see Figure 2-2 C&D for exemplar trees). However, such topologies could also be reconciled as two successive duplications with one branch death. Future analyses with additional genomes and multiple rooting strategies are needed to resolve these alternatives.

Duplications in asterids

In order to resolve duplication events within asterids, we sorted 516,751 Asterideae unigenes, from 13 asterid I species and 30 asterid II species, into tribes, identifying 70,050 asterid unigenes that were members of 1787 focal duplication tribes. These 516,751 sequences, representing more than 35,000 unigenes each from *Solanum tuberosum*, *Solanum lycopersium*, *Helianthus annuus*, and *Nicotiana tabacum*, and a smaller amount of additional data from other

related asterid species, provide sufficient data to populate most of the alignments with multiple asterid sequences, thus providing ample power for the detection of duplications in asterids. By analyzing 1787 asterid gene trees, we identified just 68 orthogroups in ML trees, 43 in MP trees, and 21 in NJ trees supporting (bootstrap >50) an asterid I + asterid II duplication (Table 2-6, AA DUP). At the 80% bootstrap support level, just 41 orthogroups (ML), 19 (MP), and 2 (NJ) indicate an asterid-wide duplication. However, at the family level, in Asteraceae, we found 223 (ML), 169 (MP), and 35 (NJ) orthogroups with good bootstrap support (>80) that suggest a large-scale duplication (Table 2-6; A2A2 DUP). In asterid I, a large number of duplications (141 (ML), 111 (MP), 32 (NJ)) occurred in Solanaceae (Nightshade family) that were not shared with Convolvulaceae (Morning-glory family) (Table 2-6, A1A1 DUP). Figure 2-4 is an exemplar tree of tribe 1200. Besides RA DUP (node 1), tribe 1200 also indicated Solanaceae wide duplication (node 2) and Asteraceae wide duplication (node 3).

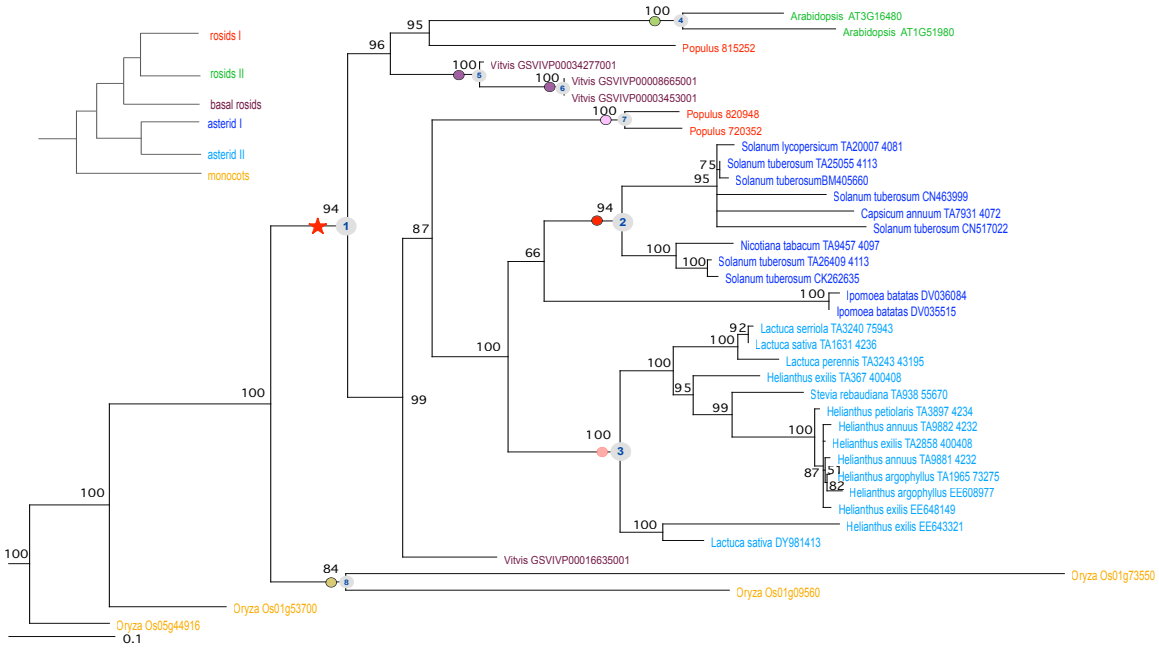


Figure 2-4. Exemplar ML topology of a tribe (Tribe 1200).

In this tribe, two major clades have survived two independent family-wide duplications (node 2: Solanaceae-wide and node 3: Asteraceae-wide). Symbols and colors same as for Figure 2-2.

Our results suggest potential WGDs in Solanaceae and Asteraceae but not across all asterids. The small number of asterid-wide gene duplications detected might be due in part to the intentionally biased sampling of tribes (e.g. tribes with AP DUP), and the lack of a fully sequenced asterid genome. On the other hand, it could truly reflect an absence of an asterid-wide genome duplication. This may be a result of a rapid radiation at the common ancestor of asterids, which would substantially reduce the time frame in which an asterid-wide WGD could have occurred; conversely, a WGD in the common ancestor of asterids could also be masked by frequent and recurrent polyploidy in the lineage. Despite the relative paucity of evidence, an increase in the number of asterid EST sequencing projects and a complete genome sequence is likely to further clarify the absence or existence of an asterid-wide WGD. Large numbers of duplications mapped to nodes 3 and 4 in Figure 2-3, do indicate large-scale duplications in Solanaceae and Asteraceae. Published K_s analyses also suggest a polyploidy event early in the evolution of Solanaceae (Schlueter et al., 2004; Cui et al., 2006).

Table 2-6. Summary of orthogroup duplications in asterids using three phylogenetic methods.

BS>=80 and BS>=50 are counts of nodes resolved with bootstrap values >=80 or >=50, respectively. Total = the number of orthogroups studied. AA DUP, A1A1 DUP, A2A2 DUP refer to nodes 6, 4, 3 of Figure 2-3, respectively, with duplications at the specified branch point.

ORTHO	ML		MP		NJ	
	BS>=80	BS>=50	BS>=80	BS>=50	BS>=80	BS>=50
TOTAL	1239	1519	913	1243	459	674
AA DUP	41	68	19	43	2	21
PERCENT	3.3%	4.5%	2.1%	3.5%	0.5%	3.1%
A1A1 DUP	141	216	111	197	32	92
PERCENT	11.4%	14.2%	12.2%	15.9%	7.0%	13.7%
A2A2 DUP	223	328	169	296	35	128
PERCENT	18.0%	21.6%	18.5%	23.8%	7.6%	19.0%

One or two rounds of duplication in the *Arabidopsis* lineage after the split of rosid I and rosid II

In 2003, Bowers *et al.* inferred a eudicot-wide "β" event and a Brassicaceae-wide "α" event (Bowers et al., 2003). However, by examining syntenic blocks through multiple aligned genomes, Tang *et al.* (Tang et al., 2008; Tang et al., 2008) proposed that the *Arabidopsis* genome has evidence of two WGDs (termed "α'" and "β'") more recent than the split of *Carica* from other Brassicales. Therefore, the previously identified (and older) Bowers *et al.* "β" should be annotated as a "γ'" event. This disagreement could be resolved by examining genetic divergences extracted from different nodes on gene family trees from our study. Two rounds of independent duplications would generate four paralogous genes, and if at least three of them have survived, the resultant topology, if observed across many gene families, could provide support for two rounds of genome-scale duplication. We investigated the topologies of trees from the 4433 initial gene family phylogenies for gene clades with at least three *Arabidopsis* sequences, using *Populus* sequences as the outgroup (Figure 2-5A, A3A4A5). Gene pairs of A4-A5 can be interpreted as being generated from the α' event and A3-A4A5 from the β' event. After removing tandem duplicates, we identified 228 (ML), 158 (MP), and 169 (NJ) clades of three or four *Arabidopsis* genes that group together with *Populus* sister (Figure 2-5B), which is likely to support two rounds of duplication. For some of these clades, three or four *Arabidopsis* genes are located on syntenic blocks with a pattern of 1 *Carica* : 4 *Arabidopsis* by examining COGE GEvo. Of the remaining groups, 1605 (ML), 1510 (MP), and 1497 (NJ) clades with only two *Arabidopsis* genes (Figure 2-5A, A6A7) were identified (Figure 2-5B). If two rounds of duplication (α' and β') occurred in the *Arabidopsis* lineage after the split of rosid I and II, the large observed number of A6A7 gene pairs should be the product of two independent WGD events. In this scenario, A6A7 duplicates should derive from two mixed sources: (1) A6A7 from the α' event with a reconciled topology of ((A6A7)(A61A71)) and gene death along the (A61A71) branch, and (2) A6A7 from the β' event

with a reconciled topology of (A6A6')(A7A7') and gene death along A6' and A7'. If only one round of duplication (α) occurred, the gene pairs of A6A7 should result from a single WGD event. Clades of A3-A4A5 could be from some other duplications (segmental, or concerted duplications (Shan et al., 2009)) other than WGD.

To determine whether the surviving A6A7 paralogs were from the α event or a mixture of α' and β', rates of nucleotide substitution per synonymous site (K_s) were used to estimate the age of gene pairs generated from each duplication event (Figure 2-5C). We first computed the mean K_s of *Arabidopsis* non-tandem gene pairs for each node with good bootstrap support (>80) indicating putative γ' (746 nodes of A1-A2), putative β'(228 nodes of A3-A4A5), and putative α' (290 nodes of A4-A5) events, and 1605 nodes of A6-A7 from ML trees (data not shown for NJ and MP). Then, we analyzed the K_s distribution within each set using mixtures of log-normals estimated by the EMMIX software (see methods). The K_s distribution of A1A2 shows a single major component (peak at K_s =2.13); however, for peaks greater than 2.0, K_s is considered to have reached saturation. The K_s distribution of A4A5 displays two peaks: 0.33 (background) and 0.78 (α' event). The K_s distribution of gene pairs of A3-A4A5, indicating the putative β' event, is estimated to have two peaks (at 1.56 and 0.81) and both contain a large proportion of pairs, which likely means that some of the A3-A4A5 gene pairs are from the β' event (peak at 1.56), and the others are from the α' event (peak at 0.81). The major component of A6A7 (peak at 0.786) is very close to the A4A5 (peak at 0.779). However, there is another significant component (peak at 1.710), which is close to the older peak of A3-A4A5 (peak at 1.56), confirming that paralogs of A6A7 were from a mixed sources of α' and β' (Figure 2-5C).

Because K_s values could be saturated at the larger values, we also examined the rate of non-synonymous substitutions (K_a) for these pairs (Figure 2-5D). Because the rate of K_a divergence is much lower than K_s divergence, the plots for K_a distributions are greatly

compressed compared to those for K_s divergences, but would potentially be able to detect a much older duplication peak if there were enough surviving genes and if the K_a divergences were not too heterogeneous across independent gene pairs (Cui et al., 2006). As described above, K_s values of A1A2 have reached saturation. The K_a distribution of A1A2 showed two peaks: 0.34 and 0.49, which indicates that 747 pairs of A1A2 were generated not only from the γ' event, but also another older one. The K_a distribution of A4A5 showed two peaks: 0.17 (α') and 0.07 (background). The K_a distribution of A6A7 showed only one major peak at 0.15, which is very close to the α' event (A4A5, 0.17). However, the K_a distribution of A3A45 has two peaks: 0.05 (background) and 0.22 (β'), of which the larger one is between the α' peak (0.17) and γ' peak (0.34). The signal supporting the β' event is detectable and significant from both K_a and K_s analysis. Therefore, the phylogenomic evidence, when integrated with genomic structural and gene pair divergences, supports the hypothesis that two rounds of WGD (α' and β') likely occurred in the *Arabidopsis* lineage after the split of rosid I and II.

Prior studies have identified either one (α, (Bowers et al., 2003)) or two rounds of WGD (α' + β', (Tang et al., 2008; Tang et al., 2008)) in the *Arabidopsis* lineage following the divergence of rosid I and rosid II. Our data show a similar pattern of duplications to that reported from Tang *et al.* (2008), but using a completely different method of analysis. Our phylogenomic-based method confirms the pattern of WGDs through time suggested by these syntenic block based approaches. The signal of unexpectedly few surviving *Arabidopsis* gene duplicates (228 clades) from a β' event should be given more attention. This might be due in part to the intentionally biased sampling of tribes (at least one *Oryza*, one *Populus* and one *Arabidopsis* gene), and the β' event mainly contributed to gene families not shared with both *Populus* and *Oryza*. The other possibility is a high gene death rate after the *Arabidopsis* β' WGD. However, small-scale (segmental or concerted duplications) duplications might also be able to generate the

current composition of *Arabidopsis* genome without β' event. This uncertainty will be addressed by sequencing another rosid lineage which shared the most recent α' event, but not the β', with Arabidopsis.

A frequent question is whether the inferred polyploidy events from Ks or structural analyses may have been due to autopolyploidy or allopolyploidy. According to the simplest classification (kihara H, 1926), autopolyploids are polyploids with multiple chromosome sets derived from a single species, which can arise through several mechanisms, including spontaneous, naturally occurring genome doubling, or following fusion of 2n gametes (unreduced gametes). Allopolyploids are polyploids with chromosomes derived from different species, which is the result of doubling of chromosome number in an F1 hybrid (kihara H, 1926). Gaut *et al.* suggested an analytical strategy that could potentially distinguish polyploid types in recent polyploids (Gaut and Doebley, 1997). However, for more ancient polyploids, although allopolyploids would possess greater initial allelic diversity than autopolyploids (all other things being equal) the genes and alleles of ancient polyploids will commonly be lost or greatly diverged from those of their initial polyploidy ancestors, making the distinction between ancient polyploidy types very challenging. As multiple new genomes are sequenced, especially for genomes of organisms derived from more recent polyploidy events, this should be a fruitful avenue for future research.

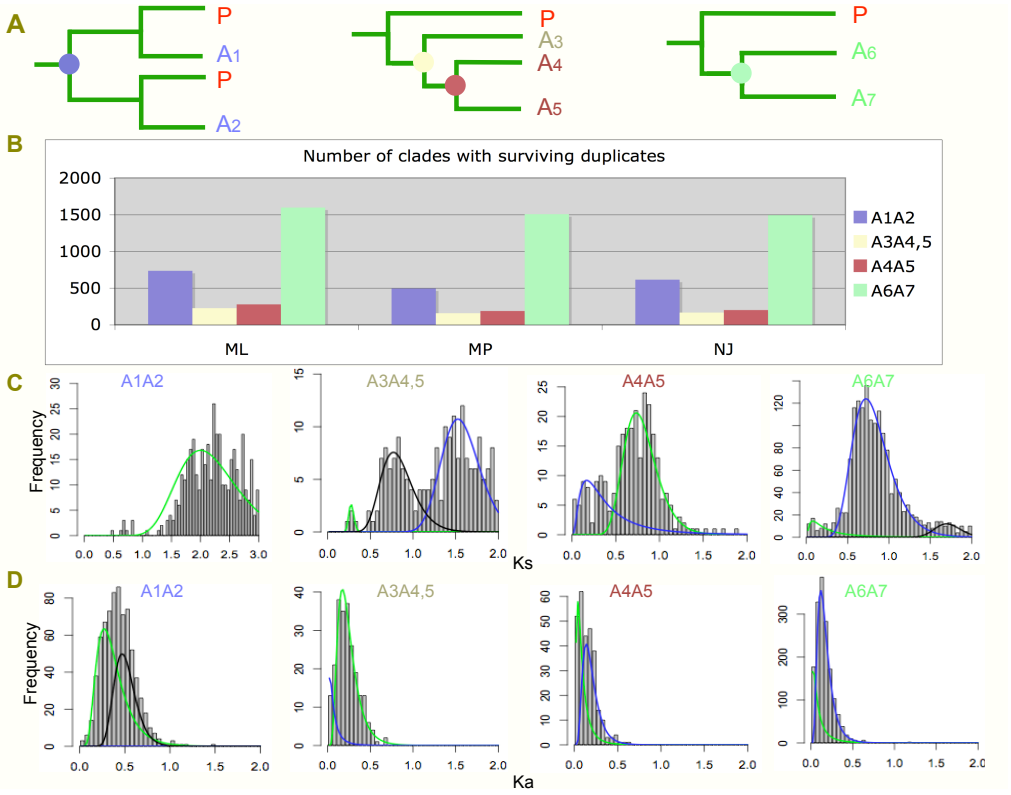


Figure 2-5. Phylogenomic analysis of gene duplications in the *Arabidopsis* lineage.

(A) Three different topologies supporting major duplication events for *Arabidopsis* lineage (tandem duplications eliminated): A1A2 –γ'; A3A4,5 –β'; A4A5 –α'; A6A7 –α'andβ'. Note: A1, A2, A3, P might represent a clade with more than one gene. (B) Number of clades with surviving *Arabidopsis* gene pairs by three different phylogenetic methods with bootstrap value greater than 80. (C) K_s distribution for duplicates from different duplication events (all from ML tree with bootstrap value greater than 80). Written as "color-mean K_s-proportion" where color is the component (curve) color, and proportion is percentage of duplication nodes assigned to the identified component. 746 nodes of A1A2: green-2.1330-1.00; 228 nodes of A3A4,5: black-0.8141-0.36, blue-1.5584-0.61, green-0.2770-0.02; 290 nodes of A4A5: blue-0.3275-0.32, green-0.7790-0.69; 1605 nodes of A6A7: blue-0.7858-0.88, black-1.7103-0.06, green-0.2121-0.06. (D) K_a distribution for duplicates from different duplication events. Written as "color-mean K_a-proportion". 746 nodes of A1A2: green-0.3359-0.60, black-0.4974-0.40; 228 nodes of A3A45: green-0.2201-0.86, blue-0.0583-0.14; 290 nodes of A4A5: blue-0.1754-0.56, green-0.0729-0.44; 1605 nodes of A6A7: blue-0.1532-0.77, green-0.0702-0.23.

Conclusions

In this study, we constructed genome-wide gene family phylogenies, reconciled them with the species tree, and inferred orthology and paralogy relationships for every pair of genes in each gene tree. These analyses enabled us to map speciation and WGD duplication events to gene trees and to evaluate them much more conclusively through the method of relative dating. With this approach, the shared rosid I- and rosid II-wide WGD can be confidently placed before the divergence of asterids and rosids. Moreover, putative large-scale duplications were detected in each of the lineages leading to Solanaceae and Asteraceae. Genetic divergences extracted from nodes on gene family trees appear to support two rounds of WGD in the history of the *Arabidopsis* lineage following the divergence of rosid I and rosid II. The phylogenomic approach is a powerful tool for studying polyploidy, and its utility will certainly increase as new EST and genomic data is rapidly being produced by next generation sequencing technologies.

Materials and methods

Gene Family Search

Putative gene families were downloaded from the PlantTribes database (Wall et al., 2008), an objectively defined database of putative protein families (Tribes) generated through the graph-based clustering algorithm MCL of six sequenced plant species: *Arabidopsis thaliana, Carica papaya* (3X draft assembly), *Medicago truncatula* (60% complete), *Oryza sativa and Populus trichocarpa* (Wall et al., 2008). Details are given in Table 2-1. We also clustered into PlantTribes the *Vitis vinifera* genome using MCL. PlantTribes gene families have been inferred using three clustering stringencies (low, medium and high) (Wall et al., 2008). Considering the balance between recovering evolutionarily complete gene family clades and obtaining good alignment profiles, we used medium stringency tribes for further analysis. Since we tried to date eudicot-wide genome duplication events, tribes with at least one *Oryza*, one *Arabidopsis* and one *Populus* sequence were selected. Preliminary analyses with *Carica* proved unsatisfactory, as *Carica* genes tended to be unstable in phylogenetic analyses, meaning that trees containing *Carica* tended to be poorly resolved with low bootstrap, and with *Carica* positioned unpredictably relative to the organismal tree (results not shown). We attributed this instability most likely to incomplete data in the 3x draft genome sequence of papaya (Ming et al., 2008). In addition, recent studies proposed two independent WGDs (α', β') occurred in the history of *Arabidopsis* lineage after the divergence with *Carica,* and no additional duplications are suspected prior to the *Arabidopsis*-*Carica* split. Therefore, *Populus* can serve as an effective outgroup for tracking WGD(s) in the recent history of *Arabidopsis*, and the *Carica* draft assembly was not essential for the analyses we performed. The *Medicago* genome is unavailable for genome-wide studies until publication. Thus, we performed our analyses with the full gene sets

drawn from *Arabidopsis*, *Populus*, *Oryza*, and *Vitis*. Unigene sets from Asterales and Solanales were downloaded from the TIGR Plant Transcript Assemblies (PTA) database (Childs et al., 2007).

Alignment

Several high throughput approaches were employed to improve overall quality of the gene family alignments and phylogenies. First, besides tribeMCL providing a preliminary quality control screen, each of the selected tribes was re-evaluated by OrthoMCL (Li et al., 2003; Tuskan et al., 2006) to identify one or more putative orthologous groups. Some highly divergent sequences were removed from the original tribes. The impact of this step is that eliminated sequences, like extinct genes, will be silent with respect to duplication events. Signals that remain of repeated duplications in genome-scale phylogeny will be conservative. All orthologous groups' amino acid alignments were generated with MUSCLE using default parameters (Edgar, 2004), and then merged together as a tribe-wide alignment by ClustalX 1.8 (Thompson et al., 2002). Corresponding DNA sequences were then forced onto the amino acid alignment using custom Perl/BioPerl scripts. Additional sorted unigene sequences for the tribe (TIGR Plant Transcript Assemblies) were aligned at the DNA level into the existing four species alignments using ClustalX 1.8 (Thompson et al., 2002). Second, these alignments were run through custom PERL scripts to identify and mask poorly aligned or non-homologous regions, which is very similar as LOW-SCORING SEGMENTS option implemented in ClustalX (Thompson et al., 2002). This program has been tested to perform well in improving large-scale gene family phylogenies. Finally, each sequence was checked, and removed from the alignment if the sequence contained less than 40% alignment coverage at the nucleotide level after masking.

Phylogenetic Analysis

Phylogenetic trees were built using Maximum Likelihood (ML), Maximum Parsimony (MP), and Neighbor-Joining (NJ) methods. Maximum likelihood (ML) analyses were conducted using RAxML version 7.0.4 (Stamatakis et al., 2005; Stamatakis, 2006) invoking a rapid bootstrap (100 replicates) analysis and search for the best scoring ML tree with the General Time Reversible model of DNA sequence evolution with gamma distributed rate heterogeneity (GTRGAMMA model, which represents an acceptable trade-off between speed and accuracy) in one single program run. MP tree searching was performed using PAUP* version 4.0b10 (Wilgenbusch and Swofford, 2003) with 10 random sequence additions and subtree-pruning-regrafting (SPR) branch-swapping with the Multrees option selected. Bootstrap analyses were performed (100 replicates) with heuristic searches using random sequence additions as above. NJ tree searching was also performed using PAUP* 4.0b10 with the Jukes-Cantor distance measure invoked and 1000 bootstrap replicates.

Scoring Gene Duplications

By carefully interpreting all of the trees, speciation and duplication events were identified and counted as orthogroups (clades) using *Oryza* genes as outgroup sequences. One orthogroup here refers to at least one *Oryza* gene, one *Arabidopsis* and one *Populus* gene grouped together. Therefore, one tribe or gene family may contain several orthogroups. Additionally, bootstrap values were taken into account, such that conflicts with the organismal topology were counted as gene duplications only if bootstrap support was greater than 80% or 50%, as indicated for alternative analyses. As result of their unstable position, *Vitis* genes were not considered to indicate a shared rosid I/II duplication event unless a duplicate copy of either *Populus* or

Arabidopsis was also observed. In other words, an orthogroup derived from a shared rosid I/II duplication should contain an observable duplication of either (AP)(A) or (AP)(P) or (AP)(AP) (A: *Arabidopsis* gene, P: *Populus* gene) with an outgroup of either *Oryza* gene or *Arabidopsis* gene or *Populus* gene.

Rate of Synonymous Substitution (K_s) Calculation

Arabidopsis duplicated pairs generated from different WGD events were identified from phylogenetic trees and assigned to the branch subtending specific nodes. For pairs of duplicates from each node, we aligned their protein sequences using muscle 3.6 (Edgar, 2004) and converted the protein alignment to DNA alignment using PAL2NAL (Suyama et al., 2006). The K_s and K_a (or dS and dN) values were calculated using a simplified version of the model of Goldman and Yang maximum likelihood method (Goldman and Yang, 1994) implemented in the codeml package of PAML (Yang, 1997), and then averaged for the specific node. The K_s frequency in each interval size of 0.05 within the range [0, 2.0] was plotted.

Finite Mixture Models of Genome Duplications

In order to explore genome duplications, a mixture model was used to treat the actual distribution of K_s between paralogs as a mixture of several component distributions in various proportions. The EMMIX software can be used to fit a mixture model of multivariate normal or t-distributed components to a given data set (http://www.maths.uq.edu.au/~gjm/emmix/emmix.html). Following Cui *et al.* (Cui et al., 2006), we modeled the log-transformed K_s distribution of paralogs. K_s values less than 0.005 were excluded to avoid fitting a component to infinity (Cui et al., 2006), and the mixed populations

were modeled with one to four components. The EM algorithm was repeated 100 times with random starting values, as well as 10 times with *k*-mean start values. The best mixture model was identified using the Bayesian Information Criterion (BIC) (Cui et al., 2006).

List of abbreviations

WGD: Whole Genome Duplication; PTA: Plant Transcript Assemblies; MP: Maximum Parsimony; NJ: Neighbor-Joining; ML: Maximum Likelihood; K_a: rate of non-synonymous substitutions per non-synonymous site; K_s: rate of synonymous substitutions per synonymous site; α, β and γ: refer to WGD events in Brassicaceae, eudicot-wide, and prior to the divergence of monocots and dicots, respectively, as proposed by Bowers *et al.* (2003); α', β': refer to WGD events in the crucifer lineage as proposed by Tang *et al.* (2008); γ': refers to a eudicot-wide WGD event as proposed by Tang *et al.* (2008).

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol **8:** 135-141

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res **13:** 137-144

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667-1678

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol **7:** R43

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438

Cannon SB, McCombie WR, Sato S, Tabata S, Denny R, Palmer L, Katari M, Young ND, Stacey G (2003) Evolution and microsynteny of the apyrase gene family in three legume genomes. Mol Genet Genomics **270:** 347-361

Cannon SB, Mitra A, Baumgarten A, Young ND, May G (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in Arabidopsis thaliana. BMC Plant Biol **4:** 10

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP (2007) The TIGR Plant Transcript Assemblies database. Nucleic Acids Res **35:** D846-851

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol **21:** 1146-1151

Conrad B, Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu Rev Genomics Hum Genet **8:** 17-35

Crane PR, Friis EM, Pedersen KR (1995) The origin and early diversification of angiosperms. Nature **374:** 27-33

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res **16:** 738-749

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

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3:** e314

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of Arabidopsis. Mol Biol Evol **23:** 469-478

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:** 113

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:** 1792-1797

Felsenstein J (1978) Cases in which Parsimony or Compatibility Methods Will be Positively Misleading. Systematic Zoology **27:** 401

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci U S A **94:** 6809-6814

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736

Goodstadt L, Ponting CP (2006) Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. PLoS Comput Biol **2:** e133

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M, Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature **431:** 946-957

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

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol **6:** 32

Johnson DA, Thomas MA (2007) The monosaccharide transporter gene family in Arabidopsis and rice: a history of duplications, adaptive evolution, and functional divergence. Mol Biol Evol **24:** 2412-2423

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428:** 617-624

kihara H OT (1926) Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. Cell and Tissue Research **4:** 475-481

Kramer EM, Jaramillo MA, Di Stilio VS (2004) Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. Genetics **166:** 1011-1023

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:** 2178-2189

Litt A, Irish VF (2003) Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. Genetics **165:** 821-833

Luo J, Yoshikawa N, Hodson MC, Hall BD (2007) Duplication and paralog sorting of RPB2 and RPB1 genes in core eudicots. Mol Phylogenet Evol **44:** 850-862

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science **290:** 1151-1155

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol **148:** 1772-1781

Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, Van de Peer Y (2005) Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A **102:** 5454-5459

Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays **27:** 937-945

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, 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 ML, 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 JK, 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 MC, 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

Ohno S (1970) Evolution by gene duplication. Springer-Verlag

Prigge MJ, Clark SE (2006) Evolution of the class III HD-Zip gene family in land plants. Evol Dev **8:** 350-361

Sampedro J, Lee Y, Carey RE, dePamphilis C, Cosgrove DJ (2005) Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. Plant J **44:** 409-419

Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci U S A **104:** 8397-8402

Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Shoemaker RC (2004) Mining EST databases to resolve evolutionary events in major crop species. Genome **47:** 868-876

Shan H, Zahn L, Guindon S, Wall PK, Kong H, Ma H, DePamphilis CW, Leebens-Mack J (2009) Evolution of plant MADS box transcription factors: evidence for shifts in selection associated with early angiosperm diversification and concerted gene duplications. Mol Biol Evol **26:** 2229-2244

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:** 2688-2690

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21:** 456-463

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res **34:** W609-612

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

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

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics **Chapter 2:** Unit 2 3

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 GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin 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, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, 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, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, 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

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. Trends Plant Sci

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, Dematte 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

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

Wall PK, Leebens-Mack J, Muller KF, Field D, Altman NS, dePamphilis CW (2008) PlantTribes: a gene and gene family resource for comparative genomics in plants. Nucleic Acids Res **36:** D970-976

Wilgenbusch JC, Swofford D (2003) Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics **Chapter 6:** Unit 6 4

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555-556

Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Soltis DE, Depamphilis CW, Ma H (2005) The evolution of the SEPALLATA subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. Genetics **169:** 2209-2223

Zhang J (2003) Evolution by gene duplication: an update. Trends Ecol Evol **18:** 292-298

Chapter 3 Ancestral polyploidy in seed plants and angiosperms

The success of angiosperms has been attributed, in part, to innovations associated with gene or whole-genome duplications (WGD; polyploidy), but evidence for hypothesized ancient genome duplications predating the divergence of monocots and eudicots remains equivocal in analyses of conserved gene order. Here we use comprehensive phylogenomic analyses of sequenced plant genomes and more than 12.6 million new EST sequences from phylogenetically pivotal lineages to elucidate two groups of ancient gene duplications – one in the common ancestor of extant seed plants and another in the common ancestor of extant angiosperms. Gene duplication events were intensely concentrated at around 319 mya and 192 mya, implicating two WGDs in ancestral lineages shortly before the diversification of extant seed plants and extant angiosperms, respectively. Significantly, these ancestral WGDs resulted in the diversification of regulatory genes important to seed and flower development, suggesting that they contributed to major innovations that ultimately contributed to the rise and eventual dominance of seed plants and angiosperms.

Background

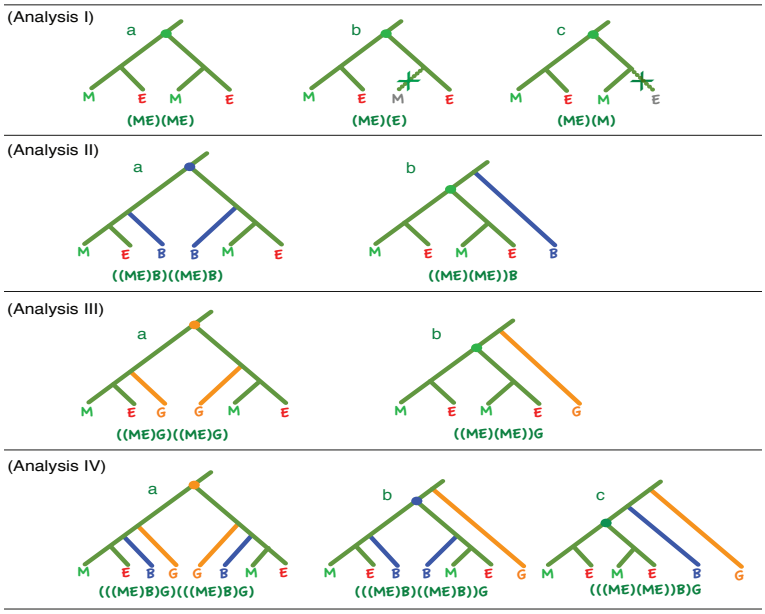
Whole-genome duplication (WGD, polyploidy) followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi, and other organisms (Ohno, 1970; Lynch, 2007; Edger and Pires, 2009), especially plants. WGDs have been hypothesized to have contributed to the origin and rapid diversification of the angiosperms (Adams and Wendel, 2005; De Bodt et al., 2005; Soltis et al., 2008; Fawcett et al., 2009), the largest group of land plants with more than 300,000 living species; significantly, most flowering plant lineages reflect one or more rounds of ancient polyploidy. For example, extensive analyses of the complete genome sequence of *Arabidopsis thaliana* provided evidence to support two

recent WGDs (named α and β) within the crucifer (Brassicaceae) lineage and one triplication event (γ) likely shared by all core eudicots (Vision et al., 2000; Bowers et al., 2003; Jaillon et al., 2007; Lyons et al., 2008; Tang et al., 2008; Barker et al., 2009), but not with monocots (Lyons et al., 2008; Tang et al., 2008). The *Populus trichocarpa* genome exhibits evidence of the core eudicot triplication as well as a more recent WGD (Tuskan et al., 2006). Two polyploidy events in monocots (ρ and σ) have been inferred to have predated the diversification of cereal grains and other grasses (Poaceae) (Tang et al., 2009). Several studies have hinted that an ancient WGD event occurred even earlier in angiosperm evolution (Vision et al., 2000; De Bodt et al., 2005; Cui et al., 2006; Soltis et al., 2008). However, the existence and timing of these ancient event(s), and their long-term impact, remain uncertain.

Several methodologies have been proposed and widely used to detect the signature of genome duplication. Identification of large syntenic blocks of genes within genomes provides strong evidence to support genome duplication (Bowers et al., 2003; Lyons et al., 2008). The timing of WGDs is inferred through cross-species genome comparisons, but extensive genome rearrangements and gene loss reduce the size of syntenic blocks over time and obscure identification of ancient pre-γ WGD(s) (Vandepoele et al., 2002; Buggs et al., 2009). Another approach is to estimate the age distribution of paralogous gene pairs, where synonymous (K_s) or nonsynonymous (K_a) site divergence is used as a proxy for the age of the duplication event (Vision et al., 2000; Blanc and Wolfe, 2004; De Bodt et al., 2005; Cui et al., 2006). However, this method may be confounded by excessive gene loss, concentration of duplicate pair estimates on more recent nodes, saturation of the rate of synonymous substitutions (K_s) between older paralogue pairs, and molecular rate heterogeneity among lineages, gene families, or even genes. For example, the β and γ genome-wide duplications inferred in analyses of syntenic blocks were not evident in a K_s plot for *Arabidopsis* paralogue pairs (Blanc and Wolfe, 2004; Tang et al.,

2008; Van de Peer et al., 2009). Therefore, both methods present challenges to inferring ancient genome duplications that may have occurred close to or well before the origin of angiosperms.

a



b

	Total	*Orthogroups w/ duplication	Type a		Type b		Type c	
			BS>=80	BS>=50	BS>=80	BS>=50	BS>=80	BS>=50
Analysis I	7470	799	234	285	142	312	91	219
Analysis II	540	345	262 (99%)	343(99%)	1(1%)	5(1%)		
Analysis III	338	229	62(51%)	147(57%)	59(49%)	110(43%)		
AnalysisIV	322	198	65(53%)	130(58%)	54(44%)	88(39%)	4(3%)	6(3%)

* Some orthogroups contain more than one type of ancient duplication as shown in Supplementary Figure 3.

Figure 3-1. Hypothetical tree topologies and corresponding summary of orthogroups consistent with ancient gene duplications prior to the split of monocots and eudicots.

(a) Analysis I: three examples of phylogenetic trees showing the patterns of retention or loss of paralogues. (Type a) both of the paralogues are retained in monocots and eudicots; (Type b) one of the paralogues was lost in monocots; (Type c) one of the paralogues was lost in eudicots. Analysis II: orthologues from basal angiosperms were added to core-orthogroups to refine the timing of ancient gene duplications in angiosperms. (Type a) gene duplication shared across all angiosperms; (Type b) gene duplication shared only by monocots and eudicots. Analysis III: orthologues from gymnosperms were added to core-orthogroups to place shared gene duplications before (Type a) and/or after (Type b) the split of extant gymnosperms and angiosperms. Analysis IV: three different topologies consistent with the timing of duplications shared by seed plants (Type a), angiosperms (Type b), and monocots + eudicots (Type c) when the expanded core-orthogroups with additional orthologues from both basal angiosperms and gymnosperms. Note: M = monocots; E = eudicots; B = basal angiosperms; G = gymnosperms. Exemplar trees in Analysis II, III, and IV illustrate expected patterns with all branches retained. Observed topologies typically had partial gene losses similar to Type b and c in Analysis I. (b) Summary of orthogroups showing different types of duplications corresponding to hypothesized topologies inferred from orthogroup trees.

Table 3-1. Summary of datasets for nine sequenced plant genomes included in this study.

Analyzed gene number is the number of genes contained in the core-orthogroups having at least one monocot and one eudicot, and one *Selaginella* and/or *Physcomitrella* sequence, which is the minimum requirement for the detection of a possible ancient duplication prior to the divergence of monocots and eudicots by a phylogenetic approach.

Species	Annotation version	Annotated genes	Analyzed genes
Arabidopsis thaliana Thale cress	TAIR version 9	27379	11669
Carica papaya Papaya	ASGPB release	25536	8713
Cucumis sativus Cucumber	BGI release	21635	9985
Populus trichocarpa Black cottonwood	JGI version 2.0	41377	15050
Vitis vinifera Grape vine	Genoscope release	30434	11020
Oryza sativa Rice	RGAP release 6.1	56979	14483
Sorghum bicolor	JGI version 1.4	34496	11258
Selaginella moellendorffii	JGI version 1.0	34697	16711
Physcomitrella patens	JGI version 1.1	35938	12551

Here, we use a phylogenomic approach to test the hypothesis that one or more ancient genome duplication(s) has occurred prior to the divergence of monocots and eudicots. By reconstructing global gene family phylogenies (Table 3-1 and 2) and mapping the duplication events onto phylogenetic trees, we determine whether the paralogous clades were duplicated before or after a given speciation event (Bowers et al., 2003; Blomme et al., 2006) (Fig. 3-1a). Although individual genes might be lost in some phylogenies, a broad picture can be drawn from simultaneous consideration of many or all gene families. In addition to available genome sequences, we also included very large, multi-tissue EST datasets derived from gymnosperm and "basal angiosperm" species (more than 12 million new sequences; Table 3-2); these

phylogenetically critical lineages increase gene sampling and provide better resolution of the timing of ancient duplications. We denote "basal angiosperms" as the earliest-branching lineages of flowering plants that arose prior to the separation of monocots and eudicots. Although they are a grade and comprise only about 10,200 species in 26 families (THE ANGIOSPERM PHYLOGENY GROUP, 2009), basal angiosperms are crucial for understanding early angiosperm evolution (Cui et al., 2006; Soltis et al., 2008; Soltis et al., 2009). In this study, they are pivotal taxa for establishing whether duplication events occurred before or after the origin of angiosperms or seed plants. Divergence time analyses were performed to test whether the timing of duplication events on these branches is localized, as would be expected if most surviving duplications were associated with genome duplication.

Table 3-2. Summary of unigene sequences of basal angiosperm and gymnosperm ESTs and unigenes included in phylogenetic study.

Gymnosperm data (except *Zamia vazquezii*) are from TIGR PTA database (http://plantta.jcvi.org/). Sequences and assemblies for basal angiosperms and *Zamia vazquezii* data (12,660,332 previously unreported ESTs) are available at the AAGP project website (Ancestral Angiosperm Genome Project) (http://ancangio.uga.edu/); data for these species will be described in detail in additional papers. Legend: # EST = total number of ESTs in the database; # Unigenes = total number of unigenes; # Included = total number of unigenes assembled in the core-orthogroups with one or more monocot + eudicot duplications.

	SPECIES (COMMON NAME)	# EST	# Unigenes	# Included
Gymnosperms	*Chamaecyparis obtusa* (Hinoki false cypress)	5830	4061	583
	Cryptomeria japonica (Japanese cedar)	16187	9098	1121
	Cycas rumphii (Cycad)	7899	4335	616
	Ginkgo biloba (Ginkgo)	5940	4178	478
	Gnetum gnemon (Melinjo)	3920	2859	195
	Picea abies (Norway spruce)	10030	5204	608
	Picea engelmannii x Picea glauca	28160	14201	1831
	Picea glauca (White spruce)	132151	49412	7782
	Picea sitchensis (Sitka spruce)	98987	25425	3047
	Pinus pinaster (Maritime pine)	13067	9166	2336
	Pinus taeda (Loblolly pine)	326641	78873	11006
	Pseudotsuga menziesii (Douglas fir)	18100	12074	291

	Taiwania cryptomerioides (Coffin tree)	1407	778	66
	Welwitschia mirabilis (Tree tumbo)	10122	6680	1408
	Zamia fischeri	8248	7374	345
	Zamia vazquezii	603139	50336	4067
Basal Angiosperms	*Aristolochia fimbriata (Dutchman's pipe)*	3828275	155371	5154
	Liriodendron tulipifera (Yellow-poplar)	2012281	141494	11582
	Nuphar advena (Yellow pond lily)	3623653	289773	27588
	Amborella trichopoda	2592984	208394	11760
	Total number of sequences	**13347021**	**1079086**	**91864**

Results and Discussion

Phylogenomic evidence for ancient gene duplications

To investigate gene duplication prior to the divergence of monocots and eudicots, we reconstructed gene families or subfamilies from species with completely sequenced genomes (Table 3-1), including two monocots (*Oryza sativa* and *Sorghum bicolor*) and five eudicots (*Arabidopsis thaliana, Carica papaya, Populus trichocarpa, Cucumis sativus,* and *Vitis vinifera*). One lycophyte (*Selaginella moellendorffii*) and one moss (*Physcomitrella patens*) were used as outgroups when dating gene duplications and potential WGDs that occurred prior to the monocot-eudicot divergence. In total, 77.03% of all protein-coding genes in the sequenced genomes were grouped in 31,433 multigene core-orthogroups. We define orthogroups as homologous gene clusters that derive from a single gene in the common ancestor of the focal taxa; orthogroups for the nine sequenced genomes are denoted here as *core-orthogroups*. Of these, 7,470 core-orthogroups contain at least one monocot, one eudicot, and one *Selaginella* and/or *Physcomitrella* sequence. These core-orthogroups were used in our investigation of duplication events predating the divergence of monocots and eudicots.

Analysis I: Construction and interpretation of phylogeny for core-orthogroups.

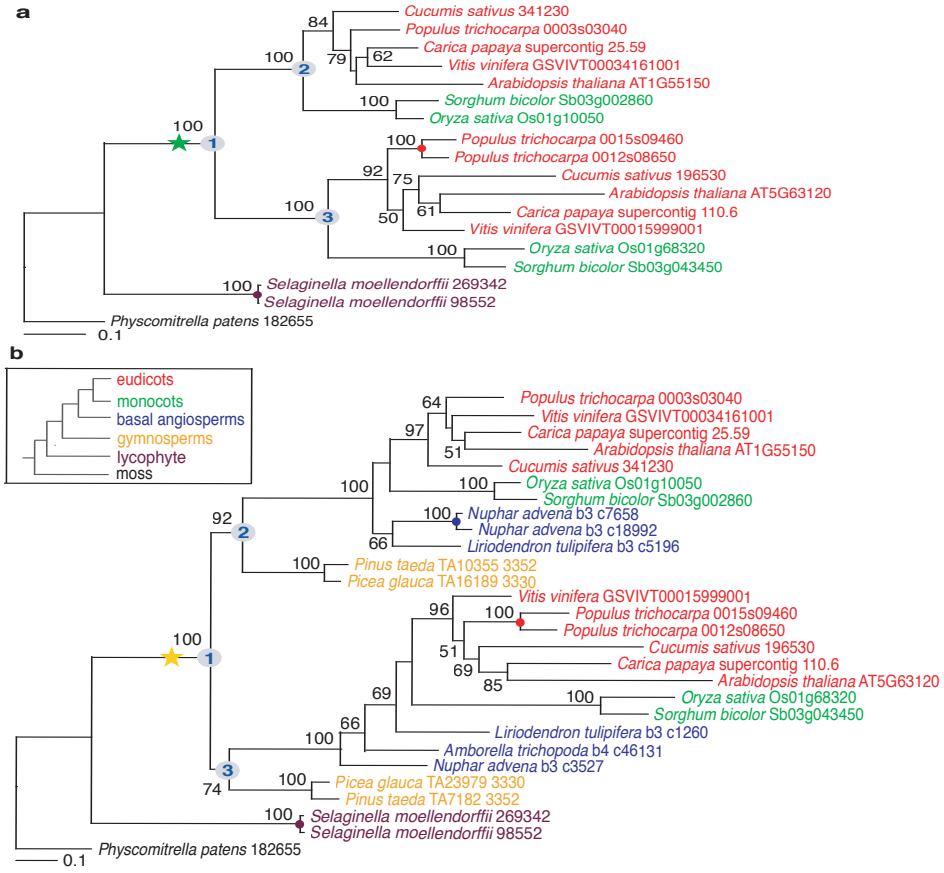


Figure 3-2. Exemplar ML phylogenies consistent with seed plant-wide duplication.

(a) RaxML topology of a core-orthogroup (Ortho 1711) where two major clades have survived the shared monocot and eudicot duplication. Since both of the duplicated clades (nodes #2 and #3) contained monocot and eudicot genes, we defined this duplication pattern as (ME)(ME). The scored BS value for this duplication is over 80%, because nodes #1 and #2 (and/or #3) have BS>80% (see "Scoring gene duplications" in Methods). (b) RaxML phylogeny of the core-orthogroup (Ortho 1711) with basal angiosperm and gymnosperm sequences added whose topology is consistent with seed plant-wide duplication. The scored BS value is over 80%, because nodes #1 and #2 have BS>80%. Legend: Green star = monocot+eudicot duplication; yellow star = seed plant duplication; colored circles = recent independent duplications; numbers = bootstrap support values.

To investigate the occurrence of ancient duplications before the divergence of monocots and eudicots, we queried Maximum Likelihood (ML) trees for each core-orthogroup for topologies indicative of shared duplications (Fig. 3-1a, Analysis I). Gene tree estimation may be

susceptible to long-branch attraction (LBA) particularly with sparse taxon sampling (i.e. gene sampling in the gene tree context) or when there is misspecification of the model of molecular evolution used for phylogenetic reconstruction (Felsenstein, 1978; Hendy and Penny, 1989), leading to erroneous conclusions of topology. For example, an orthogroup with the phylogenetic pattern ((*Oryza*, *Populus*)(*Arabidopsis*)) is consistent with a gene duplication shared by monocots and eudicots, with subsequent paralogue losses in both monocot and eudicot lineages (Fig. 3-1a, Analysis I type b). Alternatively, it is possible that the *Arabidopsis* gene was especially divergent and therefore was placed as sister to the *Oryza-Populus* pair due to LBA. Distinguishing between these alternative explanations can be facilitated by increased gene sampling to split long branches (Hendy and Penny, 1989). Moreover, inference of gene duplication may be ambiguous if all taxa are represented by a single gene in a given gene tree (as in the example above). With these considerations in mind, we filtered our gene trees, requiring that at least one of the seven core species has retained both paralogues following the inferred gene duplication event in a common monocot-eudicot ancestor. For example, the ML tree for orthogroup 1711 (DEAD box RNA helicase) contained duplicate genes in both monocots and eudicots whereas the ML tree for orthogroup 2312 (spermidine synthase) and orthogroup 396 (function unknown) showed that either one of the monocot or eudicot paralogues was lost after the divergence of monocots and eudicots (see exemplar trees in Fig. 3-2a, Fig. 3-3a and Fig. 3-4). Based on this conservative criterion, we identified a large number of core-orthogroups with shared duplication of monocots and eudicots (829 duplications in 799 core-orthogroups with bootstrap support (BS) >=50%; 474 duplications in 451 core-orthogroups with BS >=80). These duplications occurred prior to the γ triplication (which is restricted to eudicots) (Lyons et al., 2008). As expected (Lyons et al., 2008; Tang et al., 2009), many younger duplications within the sampled eudicot lineages were also observed on these trees (1146 orthogroups surviving at least one eudicot-wide triplication [the γ

event (Tang et al., 2008)]), but for this study we focused on ancient duplications that occurred prior to the divergence of monocots and eudicots.

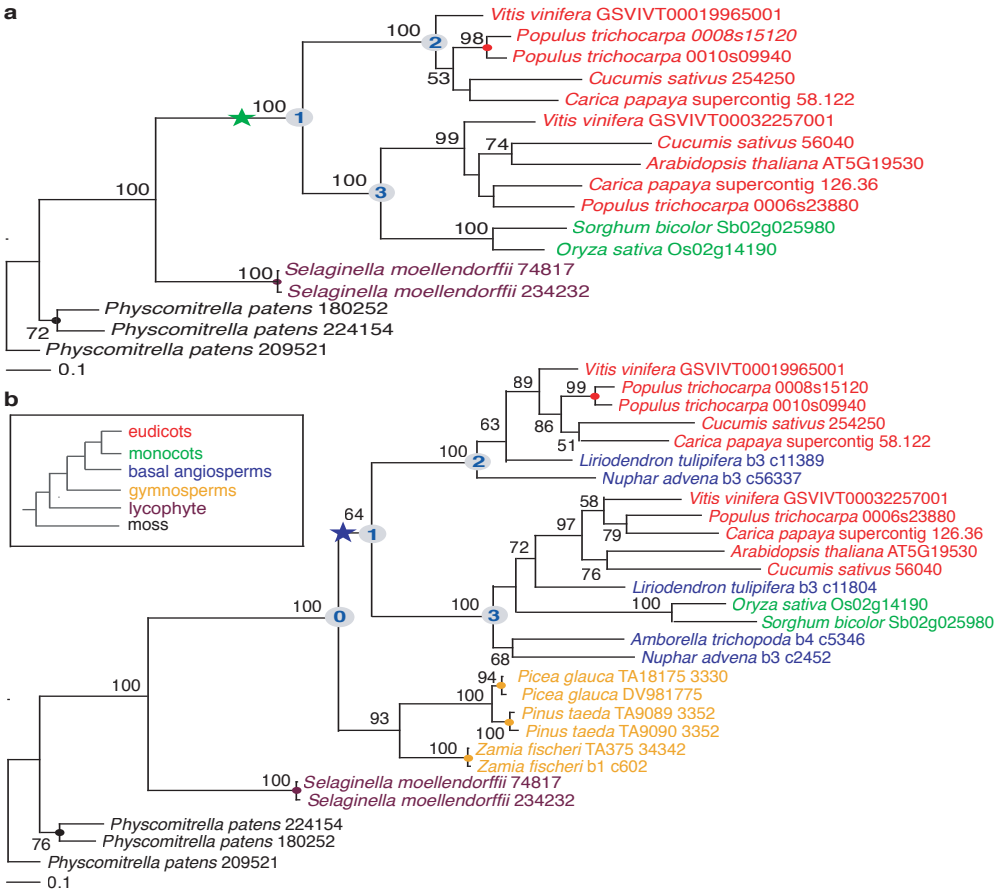


Figure 3-3. **Exemplar ML phylogenies consistent with angiosperm-wide duplication.**

(a) RaxML topology of a core-orthogroup (Ortho 2312) where two major clades have survived the shared monocot and eudicot duplication. The upper clade (node #2) only retains eudicot genes, while the lower clade (node #3) retains both monocot and eudicot genes. We defined this duplication pattern as (ME)(E). The scored BS value for this duplication is over 80% because nodes #1 and #2 (and/or #3) have BS values over 80%. (b) RaxML phylogeny of the Ortho 2312 with basal angiosperm and gymnosperm sequences added whose topology is consistent with an angiosperm-wide duplication not shared with gymnosperms. The scored BS is over 50%, because node #1 is over 50% and less than 80%. Symbols and colors same as for Figure 3-2.

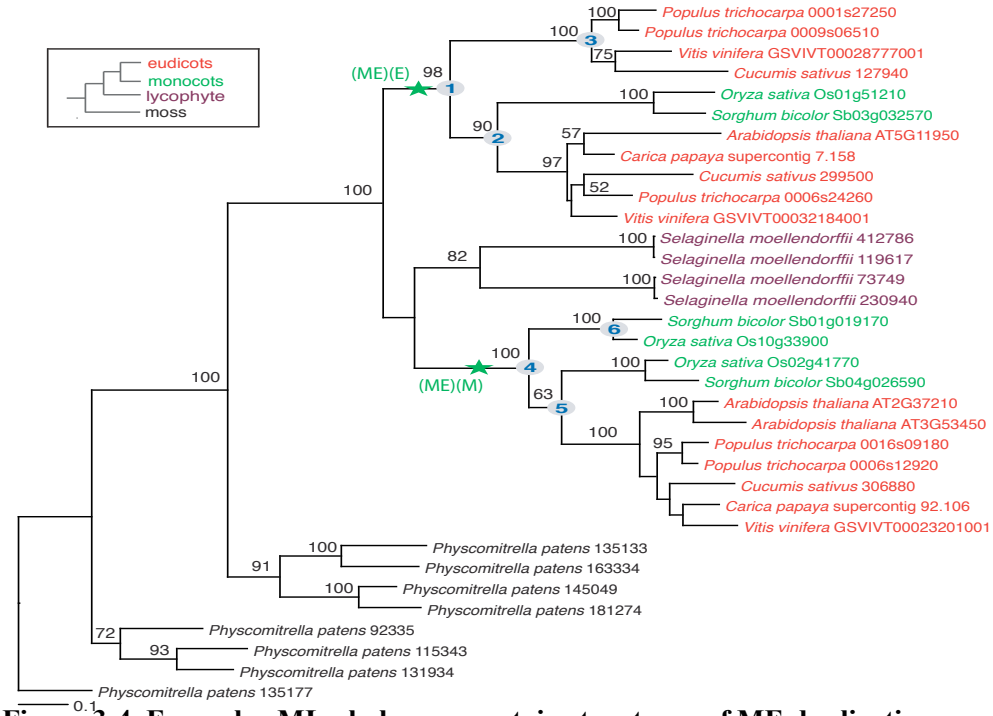


Figure 3-4. Exemplar ML phylogeny contains two types of ME duplication.

RaxML topology of a core-orthogroup (Ortho 396) with two types of shared monocot and eudicot duplications surviving. The upper part of the tree was scored as (ME)(E) with bootstrap support over 80% (Type b), since both of the BS values of node #1 and #2 were over 80%. The lower part tree was scored as (ME)(M) with bootstrap support over 50% (Type c). If one of the paralogous clades had lost all monocot or eudicot genes, the BS value of the ME clade, together with the BS of the large clade, would have been used to determine the bootstrap support level of the duplication. For the lower part of the tree, the duplication was scored BS>50%, because the BS of node #5 is 63%, even though node #4 has BS>80%. This orthogroup was counted once as Type b and once as Type c of Analysis I. Symbols and colors same as for Figure 3-2.

Increased taxon sampling in orthogroups containing basal angiosperms and gymnosperms.

Additional homologues from basal angiosperms (*Aristolochia*, *Liriodendron*, *Nuphar,* and *Amborella,* Table 3-2) and gymnosperms (*Pinus, Picea, Zamia, Cryptomeria* and others, Table 3-2) were added to 799 core-orthogroups to form expanded orthogroups (Ebersberger et al., 2009). Before re-estimating gene trees for the expanded orthogroups, we added another quality control step to remove short or highly divergent unigenes (sequences produced from assembly of

EST datasets, see Methods). After filtering, there remained 540 and 338 orthogroups with unigenes sampled from basal angiosperms and gymnosperms, respectively. Among these, 322 orthogroups contained unigenes from both basal angiosperms and gymnosperms (Fig. 3-1b).

Analysis II: Topological evidence for duplication events when basal angiosperms are considered.

For the 540 orthogroups with unigenes from basal angiosperms, the number of trees that identified a shared duplication before the origin of angiosperms but before the divergence of the monocot and eudicot sister clades (Moore et al., 2007) (Fig. 3-1a, Analysis II type a) greatly exceeded the number that identified a shared duplication after the origin of angiosperms (Fig. 3-1a, Analysis II type b). A duplication predating the diversification of basal angiosperms (ancestral angiosperm duplication) was supported by 262 (BS>=80%) or 343 (BS>=50%) orthogroups, whereas only 1 (BS>=80%) or 5 (BS>=50%) orthogroup(s) contained a gene duplication just after the origin of the angiosperm crown group (Fig. 3-1b, Analysis II). We also found only 5 orthogroups with a surviving duplication shared with some, but not all, sampled basal angiosperms. Because the duplication signal was so predominantly inclusive of all basal angiosperms, we used a single line as the placeholder for the grade of basal angiosperms in Fig. 3-1a.

K_s analysis

This finding contrasts with earlier studies based on K_s distributions of duplicated genes in basal angiosperms using much smaller numbers of ESTs (Cui et al., 2006; Soltis et al., 2009). The previous analyses had detected evidence of an ancient WGD event in several basal angiosperms, but not in *Amborella*, the sister to all other extant angiosperms (Soltis et al., 1999). However, K_s

analysis with the greatly expanded set of ESTs (2,592,984) can now detect two significant ancient duplication peaks: 1.97 and 2.76 (Fig. 3-5). Furthermore, forty-three *Amborella* unigene pairs were identified from transcriptome K_s analysis where both genes mapped to a phylogenetic tree. Gene pairs with small K_s values (<1.5) were duplicated after the divergence of *Amborella* and the rest of the angiosperms, while gene pairs with large K_s values were generated by ancient duplications in the phylogenetic trees. The results are generally consistent between the K_s and phylogenomic timing analyses. Therefore, both phylogenetic and K_s analyses of very large EST datasets provide strong evidence for a concentration of duplication events prior to the origin of angiosperms.

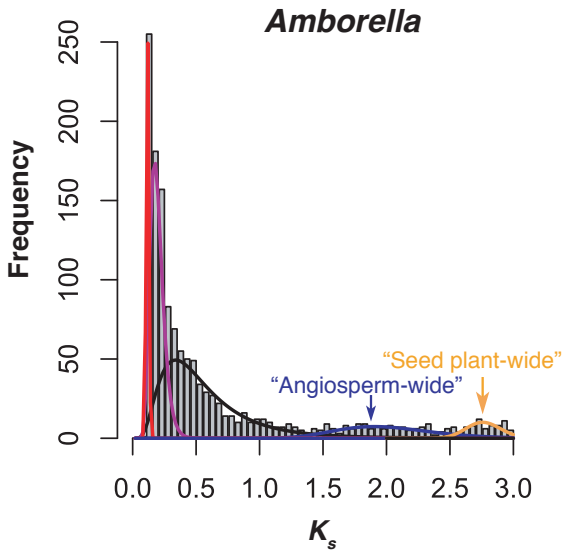


Figure 3-5. Ks distribution of 1365 paralogues in *Amborella* support ancient genome duplications.

Pairwise Ks divergences for reciprocal 'best hit' genes in Amborella EST assembly (2,592,984 ESTs). Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTN searches. Methods for sequence alignment and estimation of Ks were as reported (Cui, et al. 2006) except that only protein-coding sequences with inferred amino acid lengths >200bp were used for Ks calculations. Colored lines superimposed on Ks distribution represent significant duplication components identified by likelihood mixture model (see Methods). Graph shows "color-mean Ks-proportion" where color is the component (curve) color, and proportion is percentage of duplication nodes assigned to the identified component. Five statistically significant components: red-0.1164-0.10, purple-0.1868-0.32, black-0.4801-0.43, blue-1.9751-0.10, and yellow-2.7643-0.05.

Analyses III and IV: Topological evidence for duplication events when gymnosperms are considered.

Additional analyses of 338 orthogroups populated with unigenes of gymnosperms identified 62 (BS>=80%) or 147 (BS>=50%) trees containing a seed plant-wide gene duplication and 59 (BS>=80%) or 110 (BS>=50%) trees with a later duplication shared only by angiosperms (Fig. 3-1b, Analysis III). In addition, analyses of the 322 orthogroups expanded with orthologues from both basal angiosperms and gymnosperms also detected similar signals of the two ancient shared duplications: 65 (BS>=80%) or 130 (BS>=50%) trees showing an ancestral seed plant duplication (see exemplar tree at Figure 2b), and 54 (BS>=80%) or 88 (BS>=50%) trees supporting an ancestral angiosperm duplication (see Fig. 3-3b and Fig. 3-2b, Analysis IV).

In summary, our topological analyses of trees from thousands of orthogroups identified 799 trees with topologies suitable for testing hypotheses concerning the presence of ancient duplications. These trees provided overwhelming support for the presence of two groups of duplications, one in the common ancestor of all angiosperms and a second in the common ancestor of all seed plants. Several mechanisms could explain the concerted patterns of gene duplication revealed in the gene trees, including WGD, or multiple segmental or chromosomal duplications. The most parsimonious interpretation of the existing data is ancient WGD. We performed divergence time analyses to investigate this hypothesis further.

Ancient duplications are concentrated in time

If an hypothesized WGD is real, one would predict that the estimated dates for gene duplication events in independent gene trees will be similar. Alternatively, if the duplications are

unrelated (i.e. a collection of independent events), a uniform distribution of duplication times within the intervals between the origin of gymnosperms and angiosperms would be expected for the ancestral angiosperm duplicates or on the branch leading to seed plants for the ancestral seed plant duplicates. We calibrated 799 core-orthogroups supporting (BS>=50%) ancient duplications before the separation of monocots and eudicots from Analysis I and estimated the divergence times of 860 nodes in 774 core-orthogroups by r8s (Sanderson, 2003) (see Methods).

We then analyzed the distribution of the inferred duplication times using a Bayesian method that assigned divergence time estimates to classes specified by a mixture model (McLachlan et al., 1999). The distribution of duplication times was bimodal, with peaks at 192±2 [95% C.I.] and 319±3 mya. Dates were clustered in two relatively short time intervals, suggesting that these duplications were clearly not uniformly distributed (Fig. 3-6a). Furthermore, we also analyzed the 499 nodes with ancient duplications in 435 orthogroups with bootstrap support >=80% (Fig. 3-6b) and found a similar distribution pattern (two components: 210±4 and 321±4 mya).

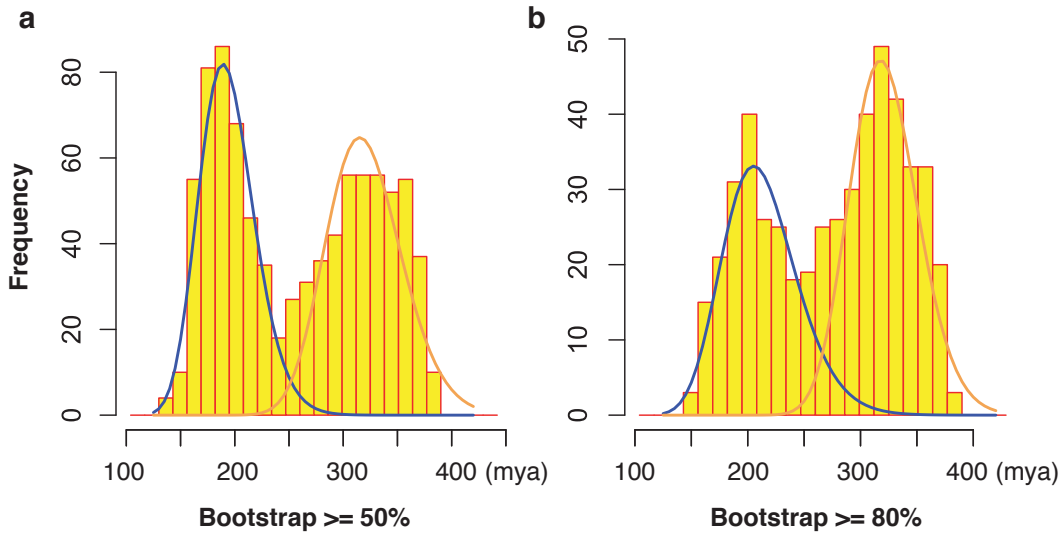


Figure 3-6. Age distribution of ancient duplications shared by monocots and eudicots.

(a) The inferred divergence times for 866 ancestral duplication nodes in 779 core-orthogroups (bootstrap support over 50%) were analyzed by EMMIX to determine whether these duplications occurred randomly over time or within some small time frame. Written as "color-mean molecular timing-proportion" where color is the component (curve) color, and proportion is percentage of duplication nodes assigned to the identified component. Two statistically significant components: blue-192(mya)-0.48 and yellow-319(mya)-0.52. (b) When we required the bootstrap support of the monocot+eudicot duplication to be over 80%, there were 504 nodes in 439 core-orthogroups for analysis of the inferred divergence times by EMMIX. Two statistically significant components were identified: blue-210(mya)-0.43 and yellow-321-0.57.

We then examined the age distribution of ancient duplications restricted only to orthogroups in Analysis III that had been populated with nearly full-length gymnosperm unigenes. Among the 338 orthogroups with inferred absolute dates, there are 110 (BS>=50%) or 59 (BS>=80%) orthogroups populated with gymnosperms that supported angiosperm-wide duplication. The distribution of duplication times inferred from these orthogroups showed one significant peak (234±9 or 236±9 mya). The most recent common ancestor of extant angiosperms occurred within the range of 130 to190 mya (Bell et al., 2005; Magallon and Sanderson, 2005; Moore et al., 2007) or possibly even earlier (Smith et al., 2010). Therefore, the identified duplication event occurred prior to the radiation of extant angiosperms, which agrees with the results from phylogenetic analysis (Fig. 3-1b, Analysis II). An additional analysis was restricted to those 147 (BS>=50%) or 62 (BS>=80%) orthogroups (Fig. 3-1b, Analysis III type a) that contained a seed-plant wide duplication based on phylogenetic analysis. The mixture model analysis identified only one significant component for the distribution of duplication times (349±3 or 347±4 mya), which was older than the ancestral node for extant seed plants (~310 mya) (Miller, 1999; Schneider et al., 2004). Thus, both molecular dating and phylogenetic analyses support another ancient genome-wide duplication shared by all extant seed plants (Fig. 3-7).

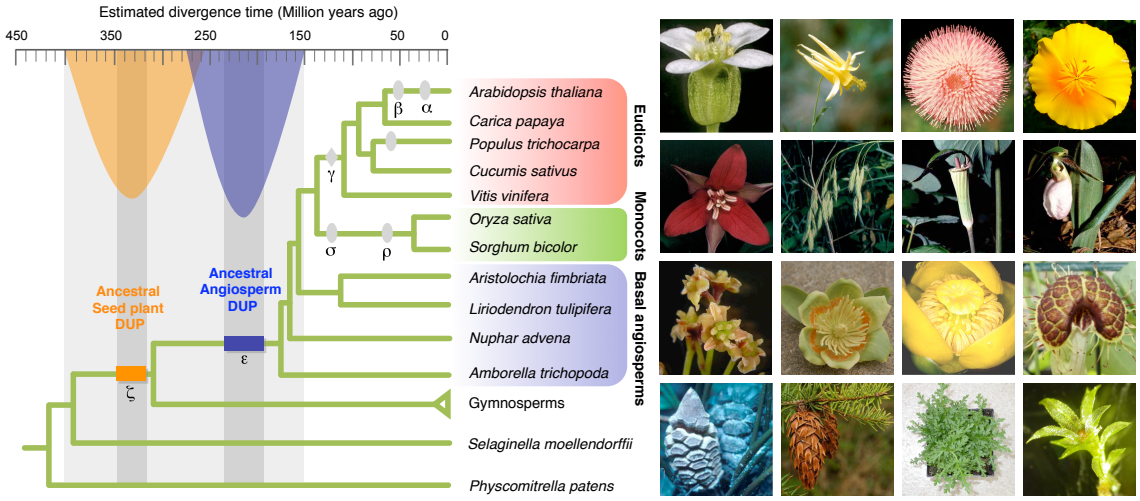


Figure 3-7. Ancestral polyploidy events in seed plants and angiosperms.

Two ancestral duplications identified by integration of phylogenomic evidence and molecular time clock for land plant evolution. Ovals indicate the generally accepted genome duplications identified in sequenced genomes (see main text for details). Diamond refers to the triplication event likely shared by all core-eudicots. Horizontal bars denote confidence regions for Ancestral Seed Plant WGD and Ancestral Angiosperm WGD, and are drawn to reflect upper and lower bounds of mean estimates from Fig. 3-6 (more orthogroups) and Analysis III (more taxa, data not shown). Photographs to right provide examples of reproductive diversity of eudicots (top row, left to right: *Arabidopsis thaliana, Aquilegia chrysantha, Cirsium pumilum, Eschscholzia californica*), monocots (second row, left to right: *Trilium erectum, Bromus kalmii, Arisaema triphyllum, Cypripedium acaule*), basal angiosperms (third row, left to right: *Amborella trichopoda, Liriodendron tulipifera, Nuphar advena, Aristolochia fimbriata*), gymnosperms (fourth row, first and second from left: *Zamia vazquezii, Pseudotsuga menziesii*), and outgroups *Selaginella mollendorfii* (vegetative; fourth row, third from left) and *Physcomitrella patens* (fourth row, right).

Synteny analysis of ancient gene duplications

A graph-based analysis of synteny(Dehal and Boore, 2005) in the *Vitis* genome was performed in order to further test the hypothesis that two ancient WGDs occurred before the divergence of monocots and eudicots and the more recent paleohexaploidy event (Jaillon et al., 2007), γ (Tang et al., 2008), that has been characterized in *Vitis* and other available eudicot genomes. A total of 2322 sets of *Vitis* genes in 571 orthogroups showing evidence of gene duplication before the monocot- eudicot divergence (Analysis I described above) were used to

test for the existence of syntenic blocks (i.e. collinear redundancy) in addition to those that that have already been described as biproducts of the γ triplication (Jaillon et al., 2007; Tang et al., 2008; Sankoff et al., 2009). Over time, rearrangements and gene loss following WGD are expected to degrade synteny between duplicated blocks in an ancient paleopolyploid genome (Jaillon et al., 2007; Lyons et al., 2008; Tang et al., 2008; Tang et al., 2009), but using the approach of Dehal and Boore (Dehal and Boore, 2005) we did find suggestive patterns of loose synteny among multiple segments that are hypothesized to have been derived from pre-γ duplication events. Gene phylogenies were used to define genes along each *Vitis* chromosome representing the pre-γ ancestral genome (Fig. 3-8) and matches between these genes were used to anchor searches for 2 or more shared genes within 200 gene windows (100 genes on either side of anchor) across the *Vitis* genome. Whereas a single pre-γ WGD would be diagnosable with up to 4-fold collinear redundancy along each chromosome (Fig. 3-8a), two pre-γ WGDs could be evidenced with a maximum of 10-fold collinear redundancy (Fig. 3-8a). However, these levels of collinearity are expected to be rare given the processes of gene fractionation following WGDs (Jaillon et al., 2007; Lyons et al., 2008; Sankoff et al., 2009; Tang et al., 2009) and structural mutations independent of WGDs. Inspection of the synteny graphs reveals that 5-fold collinear redundancy is the most common pattern in the *Vitis* genome (Fig. 3-8b), meaning that the largest fraction of genes shared another 4 paralogous regions. This result supports the phylogenomic inference that at least two pre-γ WGDs contributed to the complexity of angiosperm genomes.

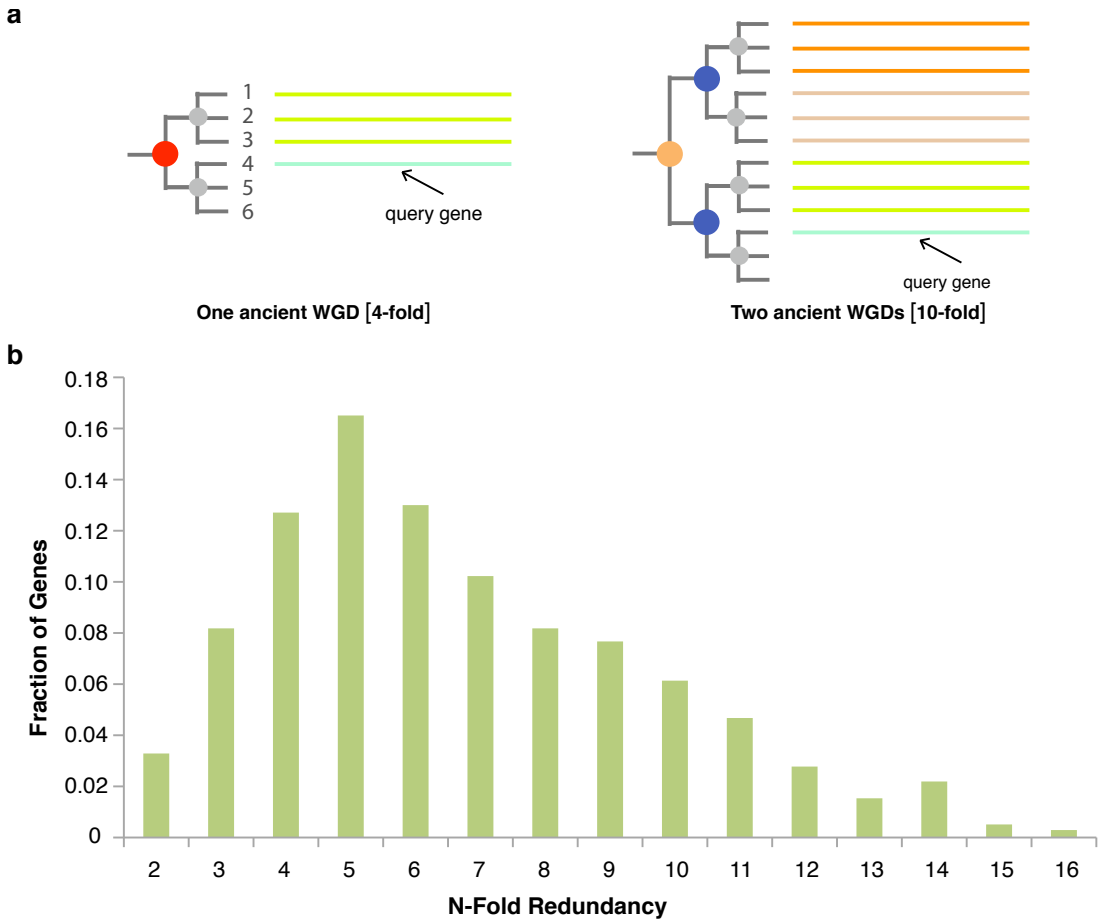


Figure 3-8. The estimate of N-fold redundancy

(a) The expected fold redundancy for hypotheses of one ancient WGD and two ancient WGDs in the history of the Vitis lineage. If one ancient WGD before the monocot-eudicot separation, six Vitis genes would be expected on the phylogenetic tree if there is no gene loss (a, left). In this case, we would identify 9 gene pairs supporting an ancient duplication before monocot-eudicot separation, which are (1,4), (1,5), (1,6), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6). Each gene would have another 3 paralogous genes on the phylogenetic tree, not including younger duplicates generated by the γ triplication. For example, gene 4 (query gene) would detect gene 1, 2, 3 as homologous genes. Therefore, the genome region where gene 4 was located would be expected to find another three paralogous regions across the Vitis genome. Therefore, one ancient WGD would lead to 4-fold redundancy (including the query). Using the same logic, two ancient WGDs would lead to 10-fold redundancy. Red filled circle refers to one ancient WGD predating the monocot-eudicot split. Yellow filled circle indicates the seed plant-wide duplication. Blue filled circles refer to the angiosperm-wide duplication. Gray filled circles denote the triplication γ event. (b) The histogram is generated by counting the redundancy across all Vitis chromosomes for each query gene. The peak at 5-fold coverage (including the query gene) means that the largest fraction of genes could detect another 4 paralogons in other regions of the genome. This is consistent with two ancient WGDs plus a more recent γ event.

Implications for plant evolution

Gene duplication provides raw genetic material for the evolution of functional novelty. WGD in ancient seed plants would have generated duplicate copies of every gene, some of which could have played crucial roles in the origin of phenotypic novelty, and ultimately in the origin and rapid diversification of the angiosperms. Although those genes retained as duplicates from the ancestral WGDs represent all functional categories, there is an overabundance of retained duplicate genes from several functional categories, including transferases and binding proteins, transcription factors, and protein kinases (Fig. 3-9). These categories are significantly enriched for orthogroups surviving ME DUP in Analysis I and for orthogroups surviving angiosperm-wide duplication and/or seed plant-wide duplication in Analysis III. These patterns are roughly consistent with post-genome duplication survivorship patterns in *Arabidopsis* (Freeling, 2009) as well as in vertebrates (Kassahn et al., 2009), and support the interpretation that the concurrent duplications observed here are products of WGD. They also suggest that the tendency for some types of gene duplications to be retained following polyploidy has been a longstanding pattern in plants for hundreds of millions of years.

One subset of duplicated genes that could have contributed to ancient seed plant and angiosperm innovations includes those that play special roles in reproductive and flower development. In this study, we identified 35 orthogroups involved in flower development pathways with at least one ancient duplication event before the divergence of monocots and eudicots (Table 3-3). For example, orthogroup 361 (*PHYTOCHROME*), which regulates the time of flowering (Devlin et al., 1998) and germination of seeds (Dechaine et al., 2009), retained duplicate genes following two putative WGDs predating the origin of angiosperms and seed plants, respectively, consistent with a published phylogeny for the *PHYTOCHROME* gene family (Mathews et al., 2003). Other published gene family phylogenies also suggested common patterns

of gene duplication, hinting at the genome-scale duplications seen here. For example, *TIR1/AFB* has experienced an ancient duplication before the diversification of extant angiosperms (Parry et al., 2009). Phylogenetic analysis of the *ZINC FINGER HOMEOBOX* (*ZHD*) family (Hu et al., 2008) and *HD-Zip III* gene family (Prigge and Clark, 2006) show duplication patterns consistent with both WGDs predating the origin of angiosperms and seed plants. Hence, these previous studies of individual genes or gene families bolster our conclusions based on a genome-wide survey of thousands of genes, and identify some of the many genes deriving from these duplications that could potentially have played important roles in seed plant and angiosperm evolution.

Many MADS-box transcription factors are important regulators of plant development, particularly as regulators of floral organ identity. Previous phylogenetic analyses of the *AGAMOUS* (*AG*), *APETALA3* (*AP3*)/*PISTILLATA* (*PI*), and *SEPALLATA* (*SEP*) MADS-box subfamilies indicated that these gene families experienced duplication prior to the eudicot-monocot divergence (Kramer et al., 1998; Kramer et al., 2004; Zahn et al., 2005; Zahn et al., 2006). The placement of basal angiosperm genes indicates that the duplication events in the *AG*, *AP3/PI*, and *SEP* subfamilies predate the diversification of extant angiosperms (Kramer et al., 2004; Stellari et al., 2004; Zahn et al., 2005; Zahn et al., 2006). These duplications are therefore consistent with WGD before the origin of the angiosperms. Furthermore, the *SEP* and *AGL6* subfamilies are sister clades formed by a duplication event that likely occurred before the split of angiosperms and gymnosperms (Zahn et al., 2005), a duplication that is possibly the same as the seed plant-wide WGD documented here. The duplication events and subsequent evolution of expression patterns and functions of these MADS-box components of the ABCE model have likely contributed to the wide spectrum of morphological diversification of flowers (Ma and dePamphilis, 2000; Zahn et al., 2005).

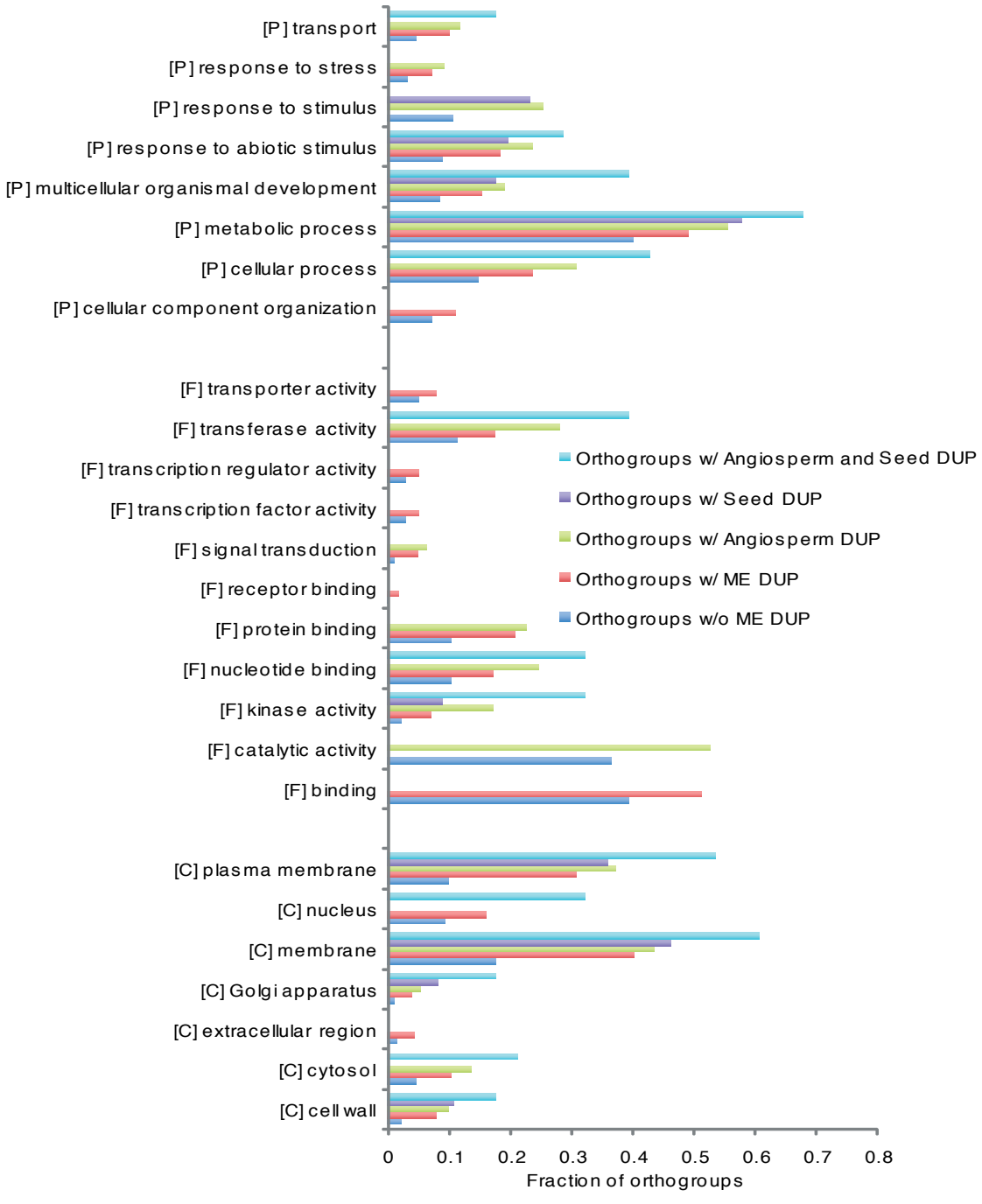


Figure 3-9. Functional categorization of orthogroups by GO annotation.

The orthogroups surviving ancient duplication (ME DUP, Angiosperm DUP, Seed plant DUP) and orthogroups without the any of these ancient duplications were categorized by GO annotation. The **X**-axis is the fraction of orthogroups mapped by the GO term and represents the abundance of the GO term. The fraction of orthogroups was calculated by the number of

orthogroups mapped to the GO term divided by the number of all orthogroups in each category. [C] means GO cellular component categorization; [F] means GO functional categorization; [P] means GO biological process categorization.

Table 3-3. Floral gene regulators surviving ancient duplications.

In this study we identified 35 orthogroups (Here just some representatives) that included genes known to regulate aspects of reproductive development in plants and containing at least one ancient gene duplication. "ME DUP" shows the number of duplications identified before the divergence of monocots and eudicots from 9-genome phylogenies. "Angio DUP" means number of angiosperm-wide duplications identified from phylogenetic trees that include basal angiosperms and gymnosperms. "Seed DUP" shows the number of seed plant-wide duplications indicated from phylogenetic trees that include basal angiosperms and gymnosperms. Numbers missing for both columns Angio DUP and Seed DUP mean the orthogroups have not been populated with unigenes of basal angiosperms and gymnosperms.

Ortho ID	Representative Gene	Annotation	ME DUP	Angio DUP	Seed DUP
34	AT1G75820	CLV1, controls shoot and floral meristem size, and contributes to establish and maintain floral meristem identity	2		
58	AT5G41170	PPR, Pentatricopeptide repeat, expressed during petal differentiation and expansion stage	1		
87	AT1G68530	CUT1, required for cuticular wax biosynthesis and pollen fertility	1		
112	AT4G04890	PDF2, encodes a homeodomain protein that is expressed in the LI layer of the vegetative, floral and inflorescence meristems	1	1	1
361	AT2G18790	PHYTOCHROME, regulates the time of flowering and seed germination	1	1	1
423	AT1G30330	ARFs, Auxin response factors, act redundantly with ARF8 to control stamen elongation and flower maturation	1		
454	AT4G32551	LEUNIG, regulates floral organ identity,gynoecium and ovule development. Negatively regulates AGAMOUS	1		1
700	AT2G42830	SHP2, SHATTERPROOF 2 (AGL5), AG, MADS box protein	1		1
1412	AT1G68050	FKF1, FLAVIN-BINDING KELCH DOMAIN F BOX PROTEIN, is clock-controlled and regulates transition to flowering	1		
1676	AT2G23380	ICU1, INCURVATA 1, required for stable repression of AG and AP3	1		

Methods

Phylogenetic analysis

The OrthoMCL method (Li et al., 2003) was used to construct a set of core-orthogroups based on protein similarity graphs. This approach has been shown to yield fewer false positives than other methods (Proost et al., 2009), which is critical for this study. If genes from outside the core-orthogroup in question (false positives) are included in the analysis, the core-orthogroup could be incorrectly scored as retaining ancient duplicates. All orthogroup amino acid alignments were generated with MUSCLE using default parameters (Edgar, 2004). The multiple sequence alignments were trimmed by removing poorly aligned regions using trimAl 1.2 with the automated1 option (Capella-Gutierrez et al., 2009). Additional sorted unigene sequences for the core-orthogroups (retrieved with HaMStR) were aligned at the amino acid level into the existing 11 species full alignments (before trimming) using ClustalX 1.8 (Thompson et al., 2002). After trimming, each unigene sequence was checked and removed from the alignment if the sequence contained less than 70% alignment coverage. Corresponding DNA sequences were then forced onto the amino acid alignment using custom Perl scripts and used for subsequent phylogenetic analysis. Maximum likelihood (ML) analyses were conducted using RAxML version 7.2.1 (Stamatakis et al., 2005; Stamatakis, 2006), invoking a rapid bootstrap (100 replicates) analysis and search for the best scoring ML tree with the General Time Reversible model of DNA sequence evolution with gamma-distributed rate heterogeneity (GTRGAMMA model, which represents an acceptable trade-off between speed and accuracy; RAxML 7.0.4 Manual) in one single program run.

Scoring gene duplications

By carefully interpreting all of the trees, duplication events were identified in rooted trees using *Physcomitrella* genes (or *Selaginella* if there were no *Physcomitrella* genes in the orthogroup) as outgroup sequences. Three relevant bootstrap values were taken into account when evaluating support for a particular duplication. For example, given a topology of (((M1E1)bootstrap1,(M2E2)bootstrap2)bootstrap3), bootstrap1 and bootstrap2 are the bootstrap values supporting the M1E1 clade and M2E2 clade, respectively, while bootstrap3 is the bootstrap value supporting the large clade including M1E1 and M2E2. A ME DUP supported by 50% (or 80%) means bootstrap3 and at least one of the bootstrap1 and bootstrap2 values >= 50% (or 80%). When basal angiosperm and/or gymnosperm genes were added, bootstrap1 and bootstrap2 were evaluated for nodes subtending ME+B (Fig. 3-1a), while bootstrap3 was evaluated for the node subtending the large clade including the angiosperm-wide or seed plant-wide duplications. For all four Analyses, we allow partial gene losses, but require that at least one of the seven sequenced angiosperm species has retained both paralogues following the inferred gene duplication event in a common ancestor of monocots and eudicots. Therefore, an example of the smallest possible gene tree with a ME DUP would be (((*Oryza*, *Vitis*)(*Vitis*))*Selaginella*). Based on these criteria, we scored each orthogroup with or without ancient duplications, and counted the total number of orthogroups supporting each hypothesis illustrated in Figure 1a.

Finite mixture models of genome duplications

In order to explore the timing of genome duplication events, the inferred distribution of divergence times was fitted to a mixture model comprising several component distributions in various proportions. The EMMIX software (McLachlan et al., 1999) can be used to fit a mixture

model of multivariate normal or t-distributed components to a given data set (http://www.maths.uq.edu.au/~gjm/emmix/emmix.html). The mixed populations were modeled with one to four components. The EM algorithm was repeated 100 times with random starting values, as well as 10 times with *k*-mean starting values. The best mixture model was identified using the Bayesian Information Criterion (BIC).

Molecular dating analyses and 95% confidence intervals

The best ML topology for the core-orthogroups or orthogroups was used for divergence time analyses. The divergence time of the two paralogous clades was estimated under the assumption of a relaxed molecular clock by applying a semi-parametric penalized likelihood (PL) approach using a truncated Newton (TN) optimization algorithm as implemented in the program r8s (Sanderson, 2003). The smoothing parameter was determined by cross-validation. We used the following dates in our estimation procedure: minimum age of 400 mya and maximum age of 450 mya for the divergence of *Physcomitrella patens (Rensing et al., 2008),* a fixed constraint age of 400 mya for the divergence of *Selaginella moellendorffii (Kenrick and Crane, 1997),* minimum age of 309 mya for crown-group seed plants (this constraint was not used in analyses reported in Figure 3-6) (Miller, 1999), minimum age of 125 mya for the split of monocots and eudicots (Doyle and Hotton, 1991), and maximum age of 125 mya for the origin of rosids (Doyle and Hotton, 1991). We required that trees pass both the cross-validation procedure and provide estimates of the age of the duplication node. The inferred divergence times were then analyzed by EMMIX. For each significant component identified by EMMIX, the 95% confidence interval of the mean was then calculated.

Rate of synonymous substitution (K_s) calculation

Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTN searches. Only protein sequences with length >200bp were used for K_s calculations. Translated sequences of unigenes generated by ESTScan were aligned using MUSCLE 3.6 (Edgar, 2004). Nucleotide sequences were then forced to fit the amino acid alignments using PAL2NAL (Suyama et al., 2006). The K_s (or *dS)* values were calculated using a simplified version of the model of the Goldman and Yang maximum likelihood method (Goldman and Yang, 1994) implemented in the codeml package of PAML(Yang, 1997). The K_s frequency in each interval size of 0.05 within the range [0, 3.0] was plotted.

GO enrichment for orthogroups with ancient duplication

GO annotations of orthogroups with early ancient duplications were compared with orthogroups that did not have such duplications, to test for enrichment of GO terms (Ashburner et al., 2000). *Arabidopsis* GO slim terms were downloaded and assigned to orthogroups directly if the orthogroup included *Arabidopsis* gene(s). Otherwise, we searched representative InterPro domains using InterProScan (Zdobnov and Apweiler, 2001). Then GO annotations were assigned to the orthogroups using InterPro2GO mapping. Subsequently, all GO annotations were mapped to GO slim categories using the map2slim script. Finally, we evaluated statistical differences in enrichment of GO slim terms using agriGO by Fisher's exact test, and the Yekutieli (FDR under dependency) multi-test adjustment method (Du et al., 2010).

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr. Opin. Plant. Biol. **8:** 135-141

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25:** 25-29

Barker MS, Vogel H, Schranz ME (2009) Paleopolyploidy in the Brassicales: analyses of the *Cleome* transcriptome elucidate the history of genome duplications in *Arabidopsis* and other Brassicales. Genome Biol. Evol. **1:** 391-399

Bell CD, Soltis DE, Soltis PS (2005) The age of the angiosperms: a molecular timescale without a clock. Evolution **59:** 1245-1258

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16:** 1667-1678

Blomme T, Vandepoele K, De Bodt S, Simillion C, Maere S, Van de Peer Y (2006) The gain and loss of genes during 600 million years of vertebrate evolution. Genome Biol. **7:** R43

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438

Buggs RJ, Doust AN, Tate JA, Koh J, Soltis K, Feltus FA, Paterson AH, Soltis PS, Soltis DE (2009) Gene loss and silencing in *Tragopogon miscellus* (Asteraceae): comparison of natural and synthetic allotetraploids. Heredity **103:** 73-81

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25:** 1972-1973

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res. **16:** 738-749

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

Dechaine JM, Gardner G, Weinig C (2009) Phytochromes differentially regulate seed germination responses to light quality and temperature cues during seed maturation. Plant Cell Environ. **32:** 1297-1309

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. **3:** e314

Devlin PF, Patel SR, Whitelam GC (1998) Phytochrome E influences internode elongation and flowering time in Arabidopsis. Plant Cell **10:** 1479-1487

Doyle JA, Hotton CL (1991) Pollen and Spores. Patterns of Diversification. Oxford, Clarendon

Du Z, Zhou X, Ling Y, Zhang Z, Su Z (2010) agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res. **38:** W64-70

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol. **9:** 157

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32:** 1792-1797

Edger PP, Pires JC (2009) Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. **17:** 699-717

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc. Natl. Acad. Sci. USA **106:** 5737-5742

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. **27:** 401

Freeling M (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annu. Rev. Plant Biol. **60:** 433-453

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725-736

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst. Zool. **38:** 297-309

Hu W, dePamphilis CW, Ma H (2008) Phylogenetic analysis of the plant-specific zinc finger-homeobox and mini zinc finger gene families. J. Integr. Plant Biol. **50:** 1031-1045

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

Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009) Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. Genome Res. **19:** 1404-1418

Kenrick P, Crane PR (1997) The origin and early evolution of plants on land. Nature **389:** 7

Kramer EM, Dorit RL, Irish VF (1998) Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics **149:** 765-783

Kramer EM, Jaramillo MA, Di Stilio VS (2004) Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. Genetics **166:** 1011-1023

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. **13:** 2178-2189

Lynch M (2007) The origins of genome architecture, Sunderland, MA

Lyons E, Pedersen B, Kane J, Alam M, Ming R, Tang H, Wang X, Bowers J, Paterson A, Lisch D, Freeling M (2008) Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. Plant Physiol. **148:** 1772-1781

Ma H, dePamphilis C (2000) The ABCs of floral evolution. Cell **101:** 5-8

Magallon SA, Sanderson MJ (2005) Angiosperm divergence times: the effect of genes, codon positions, and time constraints. Evolution **59:** 1653-1670

Mathews S, Burleigh JG, Donoghue MJ (2003) Adaptive evolution in the photosensory domain of phytochrome A in early angiosperms. Mol. Biol. Evol. **20:** 1087-1097

McLachlan G, Peel D, Basford KE, Adams P (1999) The EMMIX algorithm for the fitting of normal and t-components. J. Stat. Softw. **4**

Miller CNJ (1999) Implications of fossil conifers for the phylogenetic relationships of living families. Bot. Rev. **65:** 239-277

Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. USA **104:** 19363-19368

Ohno S (1970) Evolution by gene duplication. Springer-Verlag

Parry G, Calderon-Villalobos LI, Prigge M, Peret B, Dharmasiri S, Itoh H, Lechner E, Gray WM, Bennett M, Estelle M (2009) Complex regulation of the *TIR1/AFB* family of auxin receptors. Proc. Natl. Acad. Sci. USA **106:** 22540-22545

Prigge MJ, Clark SE (2006) Evolution of the class III HD-Zip gene family in land plants. Evol. Dev. **8:** 350-361

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K (2009) PLAZA: a comparative genomics resource to study gene and genome evolution in plants. Plant Cell **21:** 3718-3731

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin IT, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, 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 SH, 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

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19:** 301-302

Sankoff D, Zheng C, Wall PK, dePamphilis C, Leebens-Mack J, Albert VA (2009) Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. J. Comput. Biol. **16:** 1353-1367

Schneider H, Schuettpelz E, Pryer KM, Cranfill R, Magallon S, Lupia R (2004) Ferns diversified in the shadow of angiosperms. Nature **428:** 553-557

Smith SA, Beaulieu JM, Donoghue MJ (2010) An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc. Natl. Acad. Sci. USA **107:** 5897-5902

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. Am. J. Bot. **96:** 336-348

Soltis DE, Bell CD, Kim S, Soltis PS (2008) Origin and early evolution of angiosperms. Ann. NY. Acad. Sci. **1133:** 3-25

Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature **402:** 402-404

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:** 2688-2690

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21:** 456-463

Stellari GM, Jaramillo MA, Kramer EM (2004) Evolution of the *APETALA3* and *PISTILLATA* lineages of MADS-box-containing genes in the basal angiosperms. Mol. Biol. Evol. **21:** 506-519

Suyama M, Torrents D, Bork P (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. **34:** W609-612

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

Tang H, Bowers JE, Wang X, Paterson AH (2009) Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl. Acad. Sci. USA **107:** 472-477

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH (2008) 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

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics **Chapter 2:** Unit 2.3

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 GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin 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, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple JC, 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, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, 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

Van de Peer Y, Fawcett JA, Proost S, Sterck L, Vandepoele K (2009) The flowering world: a tale of duplications. Trends Plant Sci. **14:** 680-688

Vandepoele K, Simillion C, Van de Peer Y (2002) Detecting the undetectable: uncovering duplicated segments in Arabidopsis by comparison with rice. Trends Genet. **18:** 606-608

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

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl. Biosci. **13:** 555-556

Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Soltis DE, Depamphilis CW, Ma H (2005) The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. Genetics **169:** 2209-2223

Zahn LM, Leebens-Mack JH, Arrington JM, Hu Y, Landherr LL, dePamphilis CW, Becker A, Theissen G, Ma H (2006) Conservation and divergence in the *AGAMOUS* subfamily of MADS-box genes: evidence of independent sub- and neofunctionalization events. Evol. Dev. **8:** 30-45

Zdobnov EM, Apweiler R (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847-848

Chapter 4 Phylogenomic dating of the gamma polyploidy event in flowering plants

It has been widely accepted that a whole genome triplication event, referred to as gamma (γ), occurred prior to the divergence of the major rosid lineages. Although the occurrence of the γ genome triplication is well supported, the nature and actual age of the event remains uncertain in within and among species analyses of conserved syntenic blocks. To address this issue, we employed a phylogenomic approach to investigate the duplication time of *Vitis* gene sets that are located on syntenic γ blocks in the *Vitis* genome. We found that 663 *Vitis* syntenically arranged γ gene pairs were grouped into 663 putative gene families. These were aligned with gene sets from the sequenced genomes of other angiosperms. Alignments were populated with large sets of unigenes obtained from transcriptomes of various Asteridae, basal eudicots (Ranunculales), non-grass monocots, magnoliids and basal angiosperms, and then used to estimate gene family phylogenies. The overwhelming majority of well-resolved *Vitis* duplications were placed before the separation of rosids and asterids and after the split of monocots and eudicots, providing evidence for the WGD (γ) early in eudicot evolution. Further, the majority of *Vitis* gene duplications were placed after the divergence of the Ranunculales and core eudicots, supporting the γ triplication was likely restricted to core-eudicots. As this study shows, reconciliation of gene trees with a species phylogeny can elucidate the timing of major events in genome evolution, even when genome sequences are only available for a subset of species represented in the gene trees. This has major implications for the utility of comprehensive transcriptome data sets for taxa in key positions of a species phylogeny.

Background

Gene duplication provides raw genetic material for the evolution of functional novelty and is considered to be a driving force in evolution (Ohno, 1970; Adams and Wendel, 2005). Whole genome duplication (polyploidy, WGD), which involves the doubling of the entire genome, has been well documented across the tree of life including ciliates (Aury et al., 2006), fungi (Jaillon et al., 2004), flowering plants (Vision et al., 2000; Blanc et al., 2003; Bowers et al., 2003; Ming et al., 2008; Tang et al., 2008; Tang et al., 2008; Fawcett et al., 2009), and vertebrate animals (Christoffels et al., 2004; Jaillon et al., 2004; Dehal and Boore, 2005). Studies in these lineages support an association between WGD and resulting gene duplications (Blanc et al., 2003; Cui et al., 2006), functional divergence in duplicate gene pairs (Duarte et al., 2006; Johnson and Thomas, 2007), phenotypic novelty (Conrad and Antonarakis, 2007), and potentially rapid increases in species diversity (De Bodt et al., 2005; Meyer and Van de Peer, 2005).

There is growing consensus that one or more rounds of WGD events occurred early during the evolution of flowering plants (Vision et al., 2000; Bowers et al., 2003; Adams and Wendel, 2005; Ming et al., 2008; Jiao et al., 2011). By constructing ancient syntenic blocks of *Arabidopsis* and dating them phylogeneticly, it was demonstrated that three genome-wide duplication events occurred in the evolutionary history of the *Arabidopsis* lineage (Bowers et al., 2003). The oldest WGD was dated before the monocot-dicot divergence, and a second ancient WGD event was shared by most, if not all, eudicots, while a recent duplication event occurred before the radiation of Brassicales. Synteny analyses of the subsequently sequenced genomes of *Vitis vinifera* (winegrape, grapevine) (Jaillon et al., 2007) and *Carica papaya* (papaya tree) (Ming et al., 2008) provided more conclusive evidence for a somewhat different scenario in terms of the number and timing of WGDs early in the history of angiosperms. Each *Vitis/Carica* genome segment can be syntenic with up to four segments in the *Arabidopsis* genome, indicating the

presence of two WGDs in the *Arabidopsis* lineage after separation from the *Vitis/Carica* lineage (Jaillon et al., 2007; Ming et al., 2008). The more ancient one (β) might appears to have occurred around the time of the Cretaceous-Tertiary (KT) extinction (Fawcett et al., 2009). Most of the *Vitis* γ blocks are syntenic with two paralogous regions in the genome, which strongly suggested an ancient hexaploidization event (Jaillon et al., 2007; Tang et al., 2008). Available genome sequences for other core Rosid species (including papaya, *Populus*, and *Arabidopsis*) show evidence of one or more rounds of polyploidy with the most ancient event within each genome represented by triplicated syntenic blocks (Lyons et al., 2008; Ming et al., 2008; Tang et al., 2008). The most parsimonious explanation for γ, therefore would be a hexaploidization event occurred before the divergence of grapevine and core Rosids because both of them have what appears to be the remains of a triplicated genome structure (Ming et al., 2008).

Although several more eudicot genomes have been completed, the nature and exact timing of the hexaploidization event (γ) is still uncertain. For instance, it has been demonstrated that two of the three homologous regions were more fractionated, suggesting a possible mechanism for the γ event (Lyons et al., 2008). Under one proposed scenario, a genome duplication event generated a tetraploid, which then hybridized with a diploid to generate a (probably sterile) triploid. Finally, a second whole genome duplication event occurred to double the triploid genome to generate a fertile hexaploid. However, a top-down approach to the characterization of syntenic blocks indicates that three corresponding regions are generally equidistant from one another (Tang et al., 2008). Therefore, more evidence is needed to discover a more definitive mechanism for the apparent hexaploidization. Moreover, the exact timing of γ is still unclear. As described above, the γ event is readily apparent in analyses of sequenced rosids genomes. Recent comparisons of regions of the *Amborella* genome and the *Vitis* synteny blocks

indicate that the γ event occurred after the origin and early diversification of angiosperms (Zuccolo et al., 2011). In addition, comparisons of the *Vitis* synteny blocks with BAC sequences from the *Solanum* (and asterid) and *Musa* (a monocot) genomes provide weaker evidence that γ may have predated the divergence of rosids and asterid, and postdates the divergence of monocots and eudicots, respectively (Jaillon et al., 2007; Ming et al., 2008).

Phylogenomic approaches can also been used to determine the relative timing of WGDs. By mapping duplications of paralogs created by a given WGD onto phylogenetic trees, we can determine whether the paralogs were duplicated before or after a given speciation event (Bowers et al., 2003). In a recent study (Jiao et al., 2011), Jiao *et al* used a similar strategy to identify two bouts of concerted gene duplications that are hypothesized to be derived from successive genome duplications in a common ancestor of living seed plants and angiosperms. When using a phylogenomic approach, extensive rate variation among species could lead to incorrect phylogenetic inferences and then could result in the incorrect placement of duplication events (Ming et al., 2008). Gene or taxon sampling can reduce variation in branch lengths and the impact of long-branch attraction in gene tree estimates (e.g. (Leebens-Mack et al., 2005)). Therefore, one should consider possible differences in substitution rates and employ careful taxon sampling to cut potential long branches when undertaking phylogenetic analysis.

Expressed sequence tag (EST) resources produced by both traditional (Sanger) and next generation methods have grown rapidly in recent years (Childs et al., 2007; Shumway et al., 2010). In PlantGDB, very large Sanger EST datasets from multiple members of Asteraceae (e.g., *Helianthus annuus,* sunflower) and Solanaceae (e.g., *Solanum tuberosum*, potato), in particular, provide good coverage of the gene sets from the two largest asterid lineages. With advances in next generation sequencing, the comprehensive transcriptome datasets are being generated for an

expanding number of species. For example, the Ancestral Angiosperm Genome Project (AAGP) has generated large, multi-tissue cDNA datasets of magnoliids and other basal angiosperms, including *Aristolochia*, *Liriodendron*, *Nuphar* and *Amborella* ((Jiao et al., 2011); http://ancangio.uga.edu). The 1000 plant transcriptome project (oneKP or 1KP, http://www.onekp.com) is generating least three gigabases of Illumina paired end RNAseq data from 1000 plant species from green algae to angiosperms (Viridaeplantae). In this study, we draw upon these resources including an initial collection of basal eudicot species that have been very deeply sequenced by the 1KP project. Six members of Papaveraceae (*Argemone mexicana*, *Eschscholzia californica*, and four species of *Papaver*) have been targeted for especially deep sequencing, with over 12 Gb of cDNA sequence derived from 4-5 tissue-specific RNAseq libraries. Three other basal eudicots (*Podocarpus peltatum* (Podocarpaceae), *Platanus occidentalis* (Platanaceae) and *Akebia trifoliata* (Lardizabalaceae) sequenced by the OneKp project (3 Gb each), the Ancestral Angiosperm Project taxa and PlantGDB EST sets for strategically-placed species (list) were employed in phylogenomic analysis of the timing of the γ genome triplication. Assembled unigenes were sorted into gene families and analyses of gene families including *Vitis* genes located on syntenic blocks in the *Vitis* genome were performed to test alternative hypotheses for the timing of the γ event.

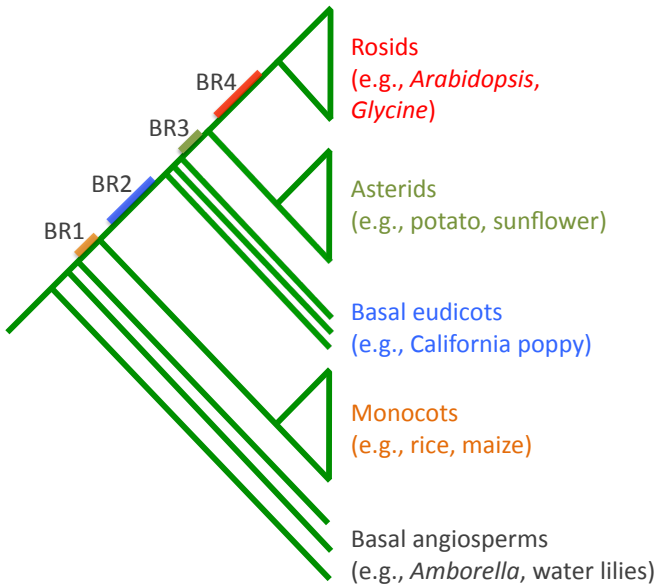


Figure 4-1. Schematic phylogenetic tree of flowering plants

BRx denote potential time points when gamma event may have occurred. BR1 – Core-angiosperm wide duplication; BR2 – Eudicot-wide duplication; BR3 – Core-eudicot wide duplication; BR4 – Rosid wide duplication.

Results and Discussion

Since the γ event was first identified in a groundbreaking phylogenomic analysis of the *Arabidopsis* genome (Bowers et al., 2003), its timing has been hypothesized to have predated the origin of angiosperms (e.g. (De Bodt et al., 2005), (Zahn et al., 2005)), the divergence of monocots and eudicots (e.g. (Chapman et al., 2006)) and the divergence of asterid and rosid eudicot clades (e.g.(Tang et al., 2008; Tang et al., 2008)) (Figure 4-1). Most recent analyses suggest that γ occurred within the eudicots, but the timing of the γ event relative to the diversification of core eudicots remains unclear. Resolving whether γ occurred just before the radiation of core eudicots or earlier, in a common ancestor of all eudicots, has implications for

our understanding of the relationship between polyploidization and speciation rates (Soltis et al., 2009).

In order to date the γ event, we downloaded *Vitis* datasets on synteny blocks from the Plant Genome Duplication Database (PGDD) (Ming et al., 2008). *Vitis* represents a basal lineage of rosid (Wang et al., 2009; Soltis et al., 2011), so to time the duplication of *Vitis* pairs phylogeneticly, homologous genes were sampled other species of rosids, asterids, basal eudicots, monocots, and basal angiosperms as outgroups. Genes were clustered into putative "orthogroups" (homologous genes that derive from a single gene in the common ancestor of the focal taxa) using OrthoMCL (Blanc et al., 2003) with eight sequenced angiosperm genomes (Table 4-1). By excluding *Vitis* pairs that are not included in same orthogroups, 783 pairs of *Vitis* genes were grouped in 671 orthogroups. These orthogroups were used in our investigation of the γ duplication event.

Table 4-1. Summary of datasets for eight sequenced plant genomes included in this study. These eight genome sequences were used to construct orthogroups, which were then populated with additional unigenes of asterids, basal eudicots, non-grass monocots, and basal angiosperms. The number of annotated genes in each genome is indicated.

Species	Annotation version	# Anotated genes
Arabidopsis thaliana Thale cress	TAIR version 9	27379
Carica papaya Papaya	ASGPB release	25536
Cucumis sativus Cucumber	BGI release	21635
Populus trichocarpa Black cottonwood	JGI version 2.0	41377
Glycine max Soybean	Phytozome version 1.0	55787
Vitis vinifera Grape vine	Genoscope release	30434

Oryza sativa Rice	RGAP release 6.1	56979
Sorghum bicolor	JGI version 1.4	34496

To establish whether the γ event was restricted to rosids or shared with asterids or even monocots, these orthogroups were then populated with unigenes of Asteridae, basal eudicots, non-grass monocots, and basal angiosperms (Table 4-2). The high rate of nucleotide substitutions and codon biases within the grasses is known to be distinct from other monocots including non-grass monocots (e.g. (Kuhl et al., 2004; Kuhl et al., 2005), so inclusion of non-grass monocots was necessary to reduce artifacts in gene tree estimation. More generally, when dealing with phylogenomic scale datasets, we have to include adequate taxon sampling to cut long branches, but avoid adding a large proportion of unigenes with low coverage. This is because inadequate taxon sampling could lead to spurious inference of phylogeny, while too many low coverage unigenes result in insufficient amounts of informative characters and then influence branch support and resolution of phylogenetic tree dramatically.

Table 4-2. Summary of unigene sequences of asterids, basal eudicots, non-grass monocots, and basal angiosperms included in phylogenetic study.

Legend: PPGP = Parasitic Plant Genome Project (http://ppgp.huck.psu.edu/); TIGR PTA = TIGR Plant Transcript Assemblies (http://plantta.jcvi.org/); AAGP = Ancestral Angiosperm Genome Project (http://ancangio.uga.edu/).

SPECIES (COMMON NAME)	Lineage	Source	# Unigenes
Panax quinquefolius	Asterid	PlantGDB	22881
Lindenbergia phillipensis	Asterid	PPGP	104904
Helianthus annuus	Asterid	TIGR PTA	44662
Solanum tuberosum	Asterid	TIGR PTA	81072
Mimulus gutatus	Asterid	PlantGDB	39577
Papaver somniferum	Basal Eudicot	1kp	252894
Papaver setigerum	Basal Eudicot	1kp	406167
Papaver rhoeas	Basal Eudicot	1kp	383426
Papaver bracteratum	Basal Eudicot	1kp	201564
Eschscholzia californica	Basal Eudicot	1kp	165260
Argemone mexicana	Basal Eudicot	1kp	148533
Akebia trifoliata	Basal Eudicot	1kp	46024
Podophyllum pelatum	Basal Eudicot	1kp	31472

Platanus occidentalis	Basal Eudicot	1kp	42373
Aquilegia formosa x Aquilegia pubescens	Basal Eudicot	PlantGDB	19615
Mesembryanthemum crystallinum	Caryophillid	PlantGDB	11317
Beta vulgaris	Caryophillid	PlantGDB	18009
Acorus americanus	Monocot	MonATOL+1kp	59453
Chamaedorea seifrizii	Monocot	MonATOL	68489
Chlorophytum rhizopendulum	Monocot	MonATOL	58766
Neoregelia sp.	Monocot	MonATOL	63269
Typha angustifolia	Monocot	MonATOL	57980
Persea americana (avocado)	Magnoliid	AAGP	132532
Aristolochia fimbriata (Dutchman's pipe)	Magnoliid	AAGP	155371
Liriodendron tulipifera (Yellow-poplar)	Magnoliid	AAGP	141494
Nuphar advena (Yellow pond lily)	Basal Angiosperm	AAGP	289773
Amborella trichopoda	Basal Angiosperm	AAGP	208394

In order to phylogenetically time a duplication event confidently, we adopted the following support based approach. Three relevant bootstrap values were taken into account when evaluating support for a particular duplication. For example, given a topology of (((clade1)bootstrap1,(clade2)bootstrap2)bootstrap3), bootstrap1 and bootstrap2 are the bootstrap values supporting clade1 (clade1 here will include one of the *Vitis* gamma duplicates) and clade2 (including the other *Vitis* duplicate), respectively, while bootstrap3 is the bootstrap value supporting the larger clade including clade1 and clade2. Duplication supported by 50% (or 80%) means bootstrap3 and at least one of the bootstrap1 and bootstrap2 values ≥ 50% (or 80%).

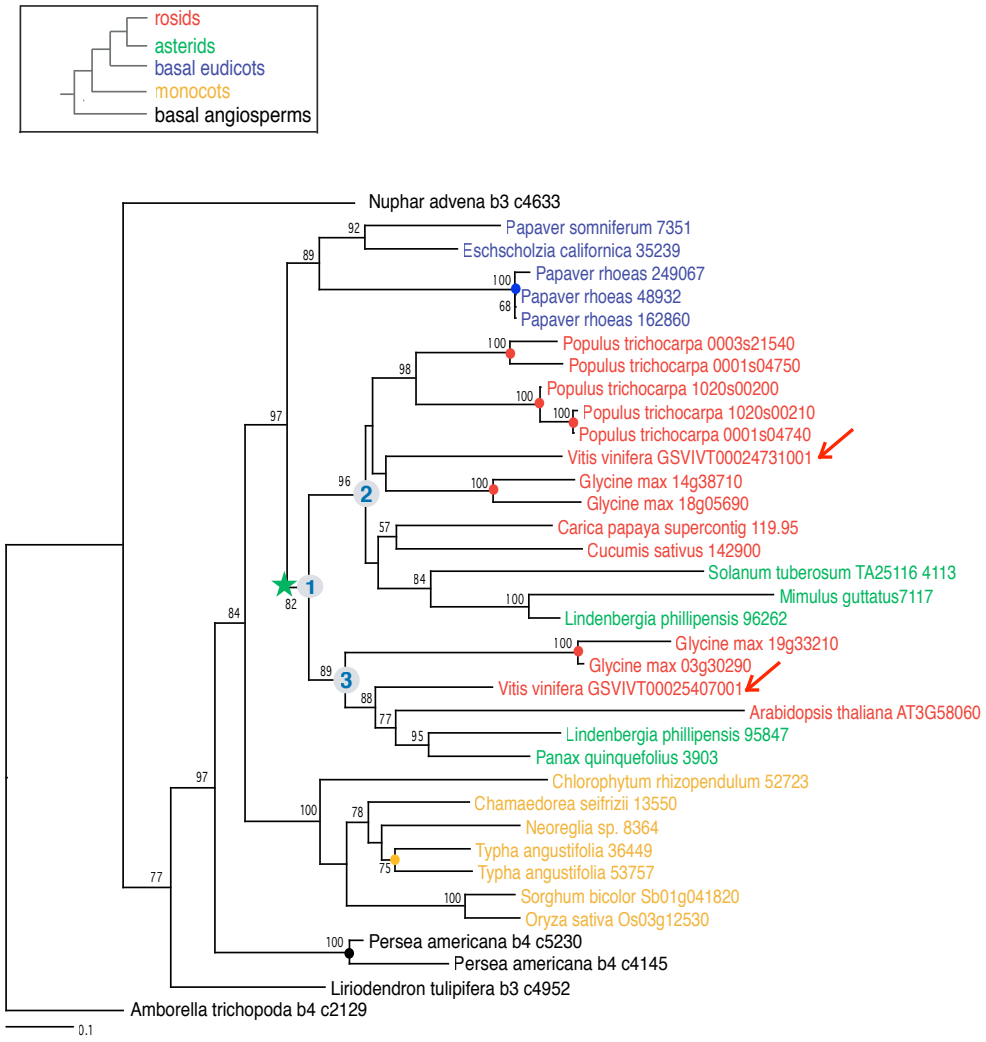


Figure 4-2. Exemplar ML phylogeny of Ortho 1202

RAxML topology of an orthogroup (Ortho 1202) indicates the gamma paralogs of *Vitis* were duplicated before the split of rosids and asterids, and after the early radiation of basal eudicots. The scored BS value for this duplication is over 80%, because nodes #1 and #2 (and/or #3) have BS >80%. Legend: Green star = core eudicot duplication; colored circles = recent independent duplications; numbers = bootstrap support values.

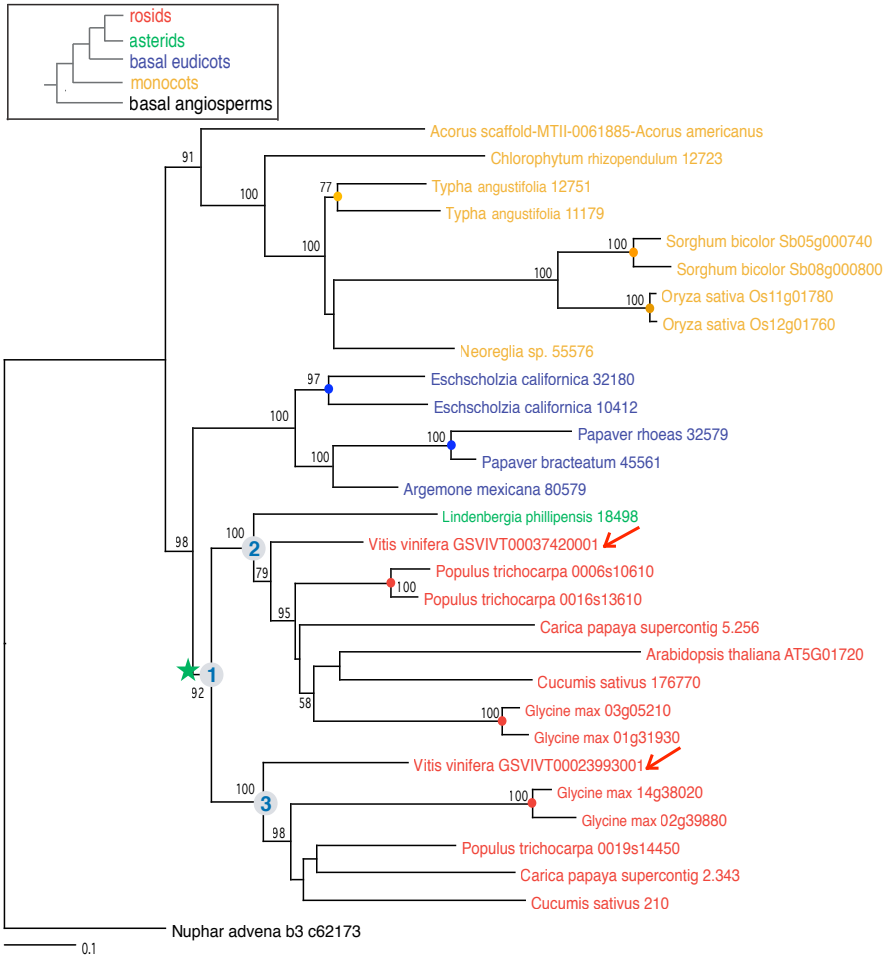


Figure 4-3. Exemplar ML phylogeny of Ortho 2606

RAxML topology of an orthogroup (Ortho 2606) indicates the gamma paralogs of *Vitis* were duplicated before the split of rosids and asterids, and after the early radition of basal eudicots. The scored BS value for this duplication is over 80%, because nodes #1 and #2 have BS >80%. Legend is same as Figure 4-2.

Homologous sequences were identified for 663 of the 671 orthogroups and were subsequently used for phylogenetic analysis. For example, orthogroup 1202 was well populated with unigenes of asterids, basal eudicots, non-grass monocots, and basal angiosperms (Figure 4-2). Two *Vitis* genes, which are located on syntenic gamma blocks, were clustered into two clades, both of which include genes from asterids and other rosids. This phylogenetic tree supports

(BS≥80%) the duplication of two *Vitis* genes before the split of rosids and asterids and after the divergence of basal eudicots. This duplication supports γ as restricted to core-eudicots (Figure 4-2). In a second example, only one asterid unigene passed the quality control steps and was clustered into orthogroup 2606. This asterid unigene was grouped into one of the duplicated clades, also supporting (BS≥80%) core-eudicot-wide hexaploidization (Figure 4-3). Only a few duplications of *Vitis* gene pairs were identified before the divergence of monocots and eudicots (1 duplication with BS≥80%; 6 duplications with BS≥50%), or restricted to rosids (3 duplications with BS≥50%). We identified 143 *Vitis* gene pairs were duplicated after the split of basal eudicots with BS≥50%, and 68 of them with BS≥80%. We also found 66 *Vitis* genes were duplicated before the separation of basal eudicots with BS≥50% and 14 with BS≥80% (Table 4-3). Therefore, our phylogenomic analysis provided overwhelming support for γ occurring before the divergence of rosids and asterids, after the split of monocots and eudicots, and most likely after the separation of basal eudicots.

Table 4-3. Phylogenetic timing of *Vitis* gamma duplications inferred from orthogroup phylogenenetic histories.

BRx designations are illustrated in Figure 4-1. BS>=80 and BS>=50 are counts of nodes resolved with bootstrap values >=80 or >=50, respectively.

ORTHO	BR1		BR2		BR3		BR4	
	BS>=80	BS>=50	BS>=80	BS>=50	BS>=80	BS>=50	BS>=80	BS>=50
Duplications	1	6	14	66	68	143	0	3
Percent	1.2%	2.8%	16.9%	30.3%	82.0%	65.6%	0%	1.4%

Phylogenetic timing of gamma duplicated *Vitis* pairs was challenging. The overwhelming majority of resolved duplications included just the “core eudicot” (Figure 4-1, BR 3 - asterids plus rosids, but not basal eudicots) groups or “all eudicots” (Figure 4-1, BR2). Other potential hypotheses for an earlier (BR1) or later (BR4) timing for the gamma event receive little or no support in the gene tree phylogenies. However, phylogenetic trees often show no resolution or

low bootstrap support for the branching nodes for *Vitis* duplicates and basal eudicots. If the gamma duplication had occurred almost anywhere along the very long stem lineage leading to eudicots, this event would have been relatively easy to resolve. The lack of resolution of branch points around the basal eudicots suggests that the dates of basal eudicots speciation events and the hexaploidization (γ) event were very close to each other. Therefore, most phylogenies were not able to strongly resolve the phylogenetic relationships among the three *Vitis* duplicates and their homologs in basal eudicots. Another possibility could be due to the nature of hexaploidization. If, as our analyses suggest, the triplication event occurred after the speciation of basal eudicots, the evolutionary rate of triplets could be different, or one of them evolve very slowly and could remain almost the same as their ancestral gene (Kellis et al., 2004). These possibilities could add significant challenges to phylogenetic resolution of events occurring at or near the branchpoints for basal versus core eudicot lineages. Despite these challenges, the majority of well-resolved gene trees support the hypothesis that the gamma triplication occurred in association with the origin and diversification of the core eudicots, after the core-eudicot lineage diverged from the Ranunculales. If this hypothesis is correct, the forthcoming *Aquilegia coerulea* genome sequence (Hodges & Derieg 2009; http://www.phytozome.net/) will become an important reference point for investigations of gene family and genome evolution following the γ hexaploidization.

Materials and methods

Phylogenetic analysis

The OrthoMCL method (Blanc et al., 2003) was used to construct sets of putative orthogroups. All orthogroup amino acid alignments were generated with MUSCLE, and then trimmed by removing poorly aligned regions by trimAl 1.2 using the heuristic automate1 option. ESTScan was used to find the best reading frame for unigenes. The predicted protein sequences were then blast against all eight sequences genomes, and sorted into orthogroups based on the best blast hit. Additional sorted unigene sequences for the orthogroups of sequenced genomes were aligned at the amino acid level into the existing full alignments (before trimming) of eight sequenced species using ClustalX 1.8. Then these large alignments were trimmed again. Each unigene sequence was checked and removed from the alignment if the sequence contained less than 70% of the total alignment length. Corresponding DNA sequences were then forced onto the amino acid alignments using custom Perl scripts and then used for subsequent phylogenetic analysis. Maximum likelihood (ML) analyses were conducted using RAxML version 7.2.1, searching for the best ML tree with the GTRGAMMA model by conducting 100 bootstrap replicates, which represents an acceptable trade-off between speed and accuracy (RAxML 7.0.4 Manual).

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol **8:** 135-141

Aury JM, Jaillon O, Duret L, Noel B, Jubin C, Porcel BM, Segurens B, Daubin V, Anthouard V, Aiach N, Arnaiz O, Billaut A, Beisson J, Blanc I, Bouhouche K, Camara F, Duharcourt S, Guigo R, Gogendeau D, Katinka M, Keller AM, Kissmehl R, Klotz C, Koll F, Le Mouel A, Lepere G, Malinsky S, Nowacki M, Nowak JK, Plattner H, Poulain J, Ruiz F, Serrano V, Zagulski M, Dessen P, Betermier M, Weissenbach J, Scarpelli C, Schachter V, Sperling L, Meyer E, Cohen J, Wincker P (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature **444:** 171-178

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res **13:** 137-144

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433-438

Chapman BA, Bowers JE, Feltus FA, Paterson AH (2006) Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci U S A **103:** 2730-2735

Childs KL, Hamilton JP, Zhu W, Ly E, Cheung F, Wu H, Rabinowicz PD, Town CD, Buell CR, Chan AP (2007) The TIGR plant transcript assemblies database. Nucleic Acids Res **35:** D846-851

Christoffels A, Koh EG, Chia JM, Brenner S, Aparicio S, Venkatesh B (2004) Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol Biol Evol **21:** 1146-1151

Conrad B, Antonarakis SE (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. Annu Rev Genomics Hum Genet **8:** 17-35

Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, Soltis PS, Carlson JE, Arumuganathan K, Barakat A, Albert VA, Ma H, dePamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res **16:** 738-749

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

Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3:** e314

Duarte JM, Cui L, Wall PK, Zhang Q, Zhang X, Leebens-Mack J, Ma H, Altman N, dePamphilis CW (2006) Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of *Arabidopsis*. Mol Biol Evol **23:** 469-478

Fawcett JA, Maere S, Van de Peer Y (2009) Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. Proc Natl Acad Sci U S A **106:** 5737-5742

Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A, Nicaud S, Jaffe D, Fisher S, Lutfalla G, Dossat C, Segurens B, Dasilva C, Salanoubat M, Levy M, Boudet N, Castellano S, Anthouard V, Jubin C, Castelli V, Katinka M, Vacherie B, Biemont C, Skalli Z, Cattolico L, Poulain J, De Berardinis V, Cruaud C, Duprat S, Brottier P, Coutanceau JP, Gouzy J, Parra G, Lardier G, Chapple C, McKernan KJ, McEwan P, Bosak S, Kellis M,

Volff JN, Guigo R, Zody MC, Mesirov J, Lindblad-Toh K, Birren B, Nusbaum C, Kahn D, Robinson-Rechavi M, Laudet V, Schachter V, Quetier F, Saurin W, Scarpelli C, Wincker P, Lander ES, Weissenbach J, Roest Crollius H (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature **431:** 946-957

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

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS, Soltis DE, Clifton SW, Schlarbaum SE, Schuster SC, Ma H, Leebens-Mack J, dePamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97-100

Johnson DA, Thomas MA (2007) The monosaccharide transporter gene family in *Arabidopsis* and rice: a history of duplications, adaptive evolution, and functional divergence. Mol Biol Evol **24:** 2412-2423

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature **428:** 617-624

Kuhl JC, Cheung F, Yuan QP, Martin W, Zewdie Y, McCallum J, Catanach A, Rutherford P, Sink KC, Jenderek M, Prince JP, Town CD, Havey MJ (2004) A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders *Asparagales* and *Poales*. Plant Cell **16:** 114-125

Kuhl JC, Havey MJ, Martin WJ, Cheung F, Yuan QP, Landherr L, Hu Y, Leebens-Mack J, Town CD, Sink KC (2005) Comparative genomic analyses in *Asparagus*. Genome **48:** 1052-1060

Leebens-Mack J, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, depamphilis CW (2005) Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol **22:** 1948-1963

Lyons E, Pedersen B, Kane J, Freeling M (2008) The value of nonmodel genomes and an expample using synmap within CoGe to dissect the hexaploidy that predates the rosids. Tropical Plant Biology **1:** 181-190

Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays **27:** 937-945

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, 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 ML, 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 JK, 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 MC, 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

Ohno S (1970) Evolution by gene duplication. Springer-Verlag

Shumway M, Cochrane G, Sugawara H (2010) Archiving next generation sequencing data. Nucleic Acids Res **38:** D870-871

Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, Depamphilis CW, Wall PK, Soltis PS (2009) Polyploidy and angiosperm diversification. Am J Bot **96:** 336-348

Soltis DE, Smith SA, Cellinese N, Wurdack KJ, Tank DC, Brockington SF, Refulio-Rodriguez NF, Walker JB, Moore MJ, Carlsward BS, Bell CD, Latvis M, Crawley S, Black C, Diouf D, Xi Z, Rushworth CA, Gitzendanner MA, Sytsma KJ, Qiu YL, Hilu KW, Davis CC, Sanderson MJ, Beaman RS, Olmstead RG, Judd WS, Donoghue MJ, Soltis PS (2011) Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot **98:** 704-730

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

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

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

Wang H, Moore MJ, Soltis PS, Bell CD, Brockington SF, Alexandre R, Davis CC, Latvis M, Manchester SR, Soltis DE (2009) Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A **106:** 3853-3858

Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, Landherr LL, Soltis DE, Depamphilis CW, Ma H (2005) The evolution of the *SEPALLATA* subfamily of MADS-box genes: a preangiosperm origin with multiple duplications throughout angiosperm history. Genetics **169:** 2209-2223

Zuccolo A, Bowers JE, Estill JC, Xiong Z, Luo M, Sebastian A, Goicoechea JL, Collura K, Yu Y, Jiao Y, Duarte J, Tang H, Ayyampalayam S, Rounsley S, Kudma D, Paterson AH, Pires JC, Chanderbali A, Soltis DE, Chamala S, Barbazuk B, Soltis PS, Albert VA, Ma H, Mandoli D, Banks J, Carlson JE, Tomkins J, Depamphilis CW, Wing RA, Leebens-Mack J (2011) A physical map for the *Amborella trichopoda* genome sheds light on the evolution of angiosperm genome structure. Genome Biol **12:** R48

Chapter 5 Conclusions

In this dissertation, we used a phylogenomic approach and constructed large-scale evolutionary trees for gene families built from a collection of genes from sequenced genomes and other large sets of ESTs from various species. By reconciling gene trees with the species tree, we inferred or dated ancient genome duplications in the history of plant. In addition, we could also use formal gene tree-species tree reconciliation, and use these to learn more about younger polyploidy event by detailed additional studying of these thousands of gene family phylogenies. We have also created a great gene tree pool that could be used to investigate the variation of evolutionary rate in different gene families. Are there any gene families evolve faster significantly than others? What is the mechanism controlling the differences? Gene balanced dosage has been proposed to explain the biased gene retention after different mode of duplications, such as “connected” proteins tending to retain after WGD. However, this conclusion is mainly based on GO enrichment analysis. Detailed analysis of evolutionary history of all members of large-scale protein complex should be done to provide stronger support for this theory.

It is difficult to determine whether WGD facilitated evolutionary transitions and diversification. However, genome structure evidence and phylogenomic study are providing more information about the number and age of ancient genome duplications. More genome sequences and new robust bioinformatic tools may unveil the correlations between polyploidy and evolutionary changes that have not been discovered yet.

VITA

Yuannian Jiao

EDUCATION

Doctor of Philosophy
- The Pennsylvania State University — 8/2006 – 12/2011
  Intercollege Graduate Degree Program in Plant Biology

Master degree
- China Agricultural University — 8/2004 – 8/2006
  Bioinformatics

Bachelor of Science
- China Agricultural University — 9/2000 – 7/2004
  Biological Sciences

CONFERENCES AND AWARDS

- ✓ PAG (Plant & Animal Genome Conferences) 2009, Poster session
- ✓ SMBE (Society for Molecular Biology and Evolution) Annual Meeting 2009
- ✓ PAG 2010 Invited talk in Polyploidy session
- ✓ ASPB (American Society of Plant Biologists) 2010, Poster session
- ✓ Braddock Research Award (2010) from PennState University
- ✓ PAG 2011 Invited talk in Polyploidy session
- ✓ Twenty-sixth Annual Graduate Exhibition of PennState University
- ✓ Penn State SMBE Symposium in honor of Masatoshi Nei's 80th Birthday

PUBLICATION

- **Yuannian Jiao**, Huanjun Guo, Xue Zhen, Lan Liu, Qunlian Zhang, Aiguang Guo, Zhen Su (2009) Discovery of *Arabidopsis GRAS* Family Genes responded to osmotic and drought stress. **Bulletin of Botany**. 44 (03): 260-268
- **Yuannian Jiao**, Norman J. Wickett, Saravanaraj Ayyampalayam, André S. Chanderbali, Lena Landherr, Paula E. Ralph, Lynn P. Tomsho, Yi Hu, Haiying Liang, Pamela S. Soltis, Douglas E. Soltis, Sandra W. Clifton, Scott E. Schlarbaum, Stephan C. Schuster, Hong Ma, Jim Leebens-Mack, Claude W. dePamphilis. (2011) Ancestral polyploidy in seed plants and angiosperms. **Nature**. 473(7345):97-100
- Haiying Liang, Saravanaraj Ayyampalayam, Norman Wickett, Abdelali Barakat, Yi Xu, Lena Landherr, Paula E. Ralph, **Yuannian Jiao**, Tao Xu, Scott E. Schlarbaum, Hong Ma, James H. Leebens-Mack, Claude W. dePamphilis (2011) Generation of a large-scale genomic resource for functional and comparative genomics in *Liriodendron tulipifera* L.. **Tree Genetics and Genomes** (*advance online*) DOI:10.1007/s11295-011-0386-2
- Andrea Zuccolo, John E. Bowers, James C. Estill, Zhiyong Xiong, Meizhong Luo, Aswathy Sebastian, Jose' Luis Goicoechea, Kristi Collura, Yeisoo Yu, **Yuannian Jiao**, Haibao Tang, Steve Rounsley, Dave Kudrna, Andrew H. Paterson, J. Chris Pires, Doug Soltis, Srikar Chamala, Brad Barbazuk, Pam Soltis, Victor A Albert, Hong Ma, Claude dePamphilis, Rod A. Wing and Jim Leebens-Mack (2011) A physical map for the *Amborella* genome sheds light on the evolution of angiosperm genome structure. **Genome Biology.** 12(5): R48