Chapter 5
Identifying the Phylogenetic Context of Whole-Genome Duplications in Plants

J. Gordon Burleigh

Abstract Although evolutionary biologists have long recognized the transformative evolutionary potential of whole-genome duplications (WGDs) in plants, identifying the precise phylogenetic location of WGDs presents many challenges. This chapter reviews some new approaches to map WGDs on a phylogeny, the first step for understanding the large-scale evolutionary and ecological consequences of WGDs in plants. Specifically, it examines approaches for using chromosome and gene copy number data, gene trees, and other genomic insights to identify the evolutionary location of WGDs. The abundance of genomic sequence data and advances in phylogenetic methods present unprecedented opportunities to place WGDs within the plant tree of life. Still, there exist few direct tests to identify and place WGDs, and analyses of complex data are often susceptible to error.

A central challenge in evolutionary biology is to determine the genetic mechanisms that generate species diversity as well as new traits, functions, and adaptations. Plant evolutionary biologists have long recognized the transformative evolutionary potential of polyploidy or whole-genome duplication (WGD) (e.g., Stebbins 1950; Grant 1963; Levin 2002; Soltis and Soltis 2000; Soltis et al. 2009). However, to study the evolutionary consequences of whole-genome duplications (WGDs) and link WGDs to phenotypic changes or diversification, the WGDs must first be placed in a phylogenetic context. Unfortunately, the rapid gene loss and genome rearrangements that frequently follow WGDs erase evidence of WGD (e.g., Wolfe 2001; Doyle et al. 2008). This process of diploidization brings about a paradox of the study of polyploidy; despite its apparent pervasiveness throughout the evolutionary history of land plants and its evolutionary importance, clear, unambiguous evidence of ancient WGDs can be remarkably difficult to locate within a phylogeny.

J. G. Burleigh (✉)
Department of Biology, University of Florida, Gainesville, FL 32611, USA
e-mail: gburleigh@ufl.edu

P. S. Soltis and D. E. Soltis (eds.), *Polyploidy and Genome Evolution*,
DOI: 10.1007/978-3-642-31442-1_5, © Springer-Verlag Berlin Heidelberg 2012

WGD in plants has been and studied for more than 100 years (e.g., Digby 1912; Winge 1917). Although sometimes cryptic, evidence of WGDs may be gleaned from such disparate sources as chromosome counts (e.g., Stebbins 1971), guard cell size (Masterson 1994), age distributions of gene copies (e.g., Blanc and Wolfe 2004), or large segmental duplications within genomes (e.g., Vision et al. 2000). Still, it is difficult to infer the phylogenetic location of the WGDs with any precision from these observations alone. To place WGDs in an evolutionary context requires a robust, and ideally well–sampled, phylogenetic hypothesis, large-scale comparative data indicating WGDs, and models and methods to map these data onto a phylogeny. Only recently have advances in phylogenetics, comparative methods, and genome sequencing made this possible on a large scale. These advances have enabled the first studies addressing some of the basic evolutionary questions about polyploidy, such as the frequency of polyploid speciation (Wood et al. 2009) or the effect of WGDs on diversification (Vamosi and Dickinson 2006; Soltis et al. 2009; Mayrose et al. 2011), in a rigorous phylogenetic framework. This chapter reviews some new approaches to map WGDs on a phylogeny, the first step for understanding the large-scale evolutionary and ecological consequences of WGDs in plants.

5.1 Chromosome Evolution

Many early surveys of polyploidy in plants used chromosome counts to estimate the percentage of polyploid species (e.g., Grant 1963, 1982; Stebbins 1938, 1950, 1971; Goldblatt 1980). This work laid the foundation for understanding the role of WGDs in plant evolution. However, lacking a phylogenetic context, it is difficult to estimate the frequency of WGD events, let alone the evolutionary placement of WGDs simply from the percentage of polyploids (although see Meyers and Levin 2006; Otto and Whitton 2002). For example, if a plant family has 100 species, 50 of which are recent polyploids, this could be the result of as few as one WGD or as many as 50 WGDs (assuming at most one WGD per lineage). Understanding the relationships among species is necessary to infer the history of WGD. Furthermore, chromosome number alone is not necessarily indicative of polyploidy. *Zea mays*, with a haploid chromosome number of 10, has had multiple WGDs in the last ~20 million years (Gaut and Doebley 1997; Gaut 2001), and *Arabidopsis thaliana*, with a haploid chromosome number of five, has experienced 3–5 WGDs since the origin of seed plants (Vision et al. 2000; Blanc et al. 2003; Bowers et al. 2003; Jiao et al. 2011). Yet few chromosome number surveys would have considered either polyploid.

Mapping chromosome numbers on a phylogenetic hypothesis can help reveal the frequency and evolutionary placement of WGD events. Informal phylogenetic observations were first used to deduce ancestral chromosome numbers. For example, numerous studies surmised a low base chromosome number for angiosperms based on the chromosome counts of the “basal” angiosperm lineages (e.g., Ehrendorfer et al. 1968; Stebbins 1971; Walker 1972; Raven 1975). These observations implied a history of WGD near the root of angiosperms, although they could not map this precisely; indeed, the relationships among these lineages were unknown. With the growth of phylogenetic methods and data, more formal maximum parsimony approaches were used to reconstruct ancestral chromosome number on the inner nodes of phylogenetic trees (e.g., Stace et al. 1997; Schultheis 2001; Mishima et al. 2002; Guggisberg et al. 2006; Hipp et al. 2007). To do this, chromosome number can be treated as a discrete variable, and the ancestral states can be reconstructed using linear or squared change parsimony. If the ancestral states are far higher than the base chromosome number, then that ancestral node may represent a polyploid. It is possible to construct elaborate chromosome number transition matrices for parsimony analyses, for example, allowing chromosome doubling as well as single chromosome changes, or to weight different changes, like down-weighting chromosome losses, but these analyses are rarely performed. In any case, parsimony reconstructions often have difficulty accounting for multiple transitions on a single branch or quantifying uncertainty in ancestral state reconstructions.

More recently, probabilistic models of chromosome number evolution have been developed (Meyers and Levin 2006; Mayrose et al. 2010). In a simple formulation, the chromosome models allow transitions that add a chromosome, remove a chromosome, or double the number of chromosomes (Mayrose et al. 2010). Thus, although the transition matrix among chromosome states (chromosome numbers) may be extremely large, the evolutionary process can be modeled with only a few parameters. The performance of these models has not been characterized in detail; however, they appear to infer more ancient WGDs than parsimony methods (Wood et al. 2009) and may also provide quite different ancestral state reconstructions of chromosome number (Cusimano et al. 2012). In the future, these models may link chromosome evolution to diversification rates or phenotypes related to the frequency of WGDs to obtain even more accurate estimates of WGDs in a phylogeny.

Studies of plant chromosome numbers have provided a wealth of insight into polyploidy in plants and have contributed substantially to canonical views of plant speciation and evolution. Yet chromosome number is a sort of summary statistic for WGD, a simple observation that is meant to represent a complex, large-scale genomic change, and chromosome number alone may not be sufficient to detect WGDs. A small chromosome number, as in *Arabidopsis thaliana*, does not necessarily imply the absence of historical WGDs, and high chromosome numbers are not necessarily evidence of WGDs. Without additional cytological or genetic data, it is impossible to distinguish between a WGD and increasing dysploidy, a change in the chromosome number that is not associated with a change in the amount of genetic material, based solely on chromosome counts.

Also, as with any phenotype, there are limitations and biases associated with ancestral state reconstruction (e.g., Schluter et al. 1997; Ané 2008). Often reconstruction is most difficult for characters with high rates of evolution or high degrees of homoplasy. Chromosome numbers may be unstable following a WGD (see Lim et al. 2008; Chester et al. 2012) and can decrease quickly following a WGD. Thus, modeling approaches likely will have difficulty for identifying ancient WGDs. In fact, it appears that the frequency of chromosome loss and diploidization was not always appreciated in studies that only examined chromosome numbers, and even with large-scale genomic data, the mechanisms for rapid chromosome loss are not clear (Doyle et al. 2008). This lack of appreciation for the lability of chromosome numbers may have contributed to the failure to detect, or even surmise, the extent of ancient WGDs, and also may have encouraged the idea that WGDs were evolutionary dead ends (e.g., Stebbins 1950).

Despite the limitations of chromosome numbers alone, data are available for many thousands of plant species, for example on the online Index to Plant Chromosome Numbers (IPCN) database (http://www.tropicos.org/Project/IPCN). Thus, until large-scale genomic sequence data sets become available for thousands of phylogenetically diverse taxa, chromosome number may provide the best opportunity to identify putative WGD events, especially recent events, with phylogenetic precision and to examine the macroevolutionary consequences of WGD throughout the history of all plants.

5.2 Gene Copy Models

With the availability of large-scale genomic data from an increasing number of plant species, estimates of copy numbers for gene families are increasingly available for many plant species. Since WGD should change not only the chromosome numbers but also the copy numbers for all gene families, gene family copy number should provide more data to estimate a WGD than simply a single chromosome number. Like chromosome number evolution, ancestral gene copy numbers can be reconstructed using parsimony methods (Snel et al. 2002; Kunin and Ouzounis 2003; Mirkin et al. 2003; Csűrös 2010; Ames et al. 2012; Librado et al. 2012). The parsimony models can be implemented in numerous ways, including weighting gains and losses differently. It is not necessarily easy to find evidence of WGDs based on the number of gene gains or losses on a branch in the species tree, but we might expect WGDs will result in far more gains, and subsequently losses, per unit time than are found on other branches. Hahn et al. (2005) developed a stochastic birth and death model that assumes a homogeneous process of duplication and loss throughout the species tree. The ML implementation of this model in CAFÉ

(De Bie et al. 2006), as well as a similar Bayesian approach (Liu et al. 2011), estimates gene family gain and loss rates across the tree and can identify anomalous gene families and branches on the tree. These branches may reflect the effects of WGDs. More complex models that account for heterogeneity in the rates of duplications and losses across lineages, and in some cases also allow gains of genes or gene families by lateral transfer, also have been proposed (Iwasaki and Takagi 2007; Csurös 2010; Ames et al. 2012; Librado et al. 2012).

In spite of much recent work on developing models of gene family copy number, all of the gene copy models assume that gene duplications or gene gains are independent. Thus, a WGD in a plant might be viewed as 20,000 gene duplications rather than a single duplication event. Consequently, while an increased duplication rate on a branch or increased loss rates on subsequent branches may suggest a WGD, there is no definitive test of WGD, and it may be difficult to distinguish WGD from simply an elevated duplication rate or a large-scale duplication. One approach may be to create a model that could estimate a rate of doubling for all gene family numbers. This transition matrix could be applied to different branches to test for either a rate of WGD greater than zero or different rates between clades or branches.

Although gene copy number provides a more detailed assessment of genomic content than chromosome numbers, inferring the histories of gene copy number and chromosome number have similar limitations. For example, gene copies appear to be rapidly silenced and lost immediately following a WGD (e.g., Tate et al. 2006; Buggs et al. 2009, 2012; see Chap. 14, this volume), which may quickly obscure the evidence for WGDs. However, simply obtaining accurate estimates of gene copy number for extant taxa may be a challenge. Without complete genome sequencing, it can be difficult to distinguish a gene loss from a failure to sample a gene. In fact, the lack of complete sequencing across a broad range of plant species may explain the lack of studies of gene copy number evolution in plants. Even with complete genome sequences, estimates of gene copy number depend on the vagaries of the extremely complex genome annotation and gene family circumscription problems. Furthermore, there are high levels of intraspecific variation in gene copy number in some plants (e.g., Springer et al. 2009; Debolt 2010; Zheng et al. 2011). It is possible to account for uncertainty in the gene copy numbers or incomplete sampling in a likelihood model, although such approaches have not been implemented.

Still, gene copy number does not always provide direct evidence for the location of gene or WGDs. For example, take the case in Fig. 5.1, in which an outgroup has a single gene copy, and two sister taxa each have two gene copies. The parsimonious explanation for these data would be that a gene duplication preceded the most recent common ancestor of the sister taxa, although it is possible that there were independent duplications in each sister lineage. In this case, the gene topologies can provide much more insight into the history of duplication than simply looking at copy number and can easily distinguish between the two duplication scenarios in Fig. 5.1. The additional information from evolutionary history of the genes can further help identify the placement of historical duplications and WGDs.

5.3 Gene Tree Reconciliation

The general problem of gene tree reconciliation is based on the observation that population-level processes, such as coalescence (lineage sorting), as well as evolutionary events, such as gene duplications and loss, recombination,

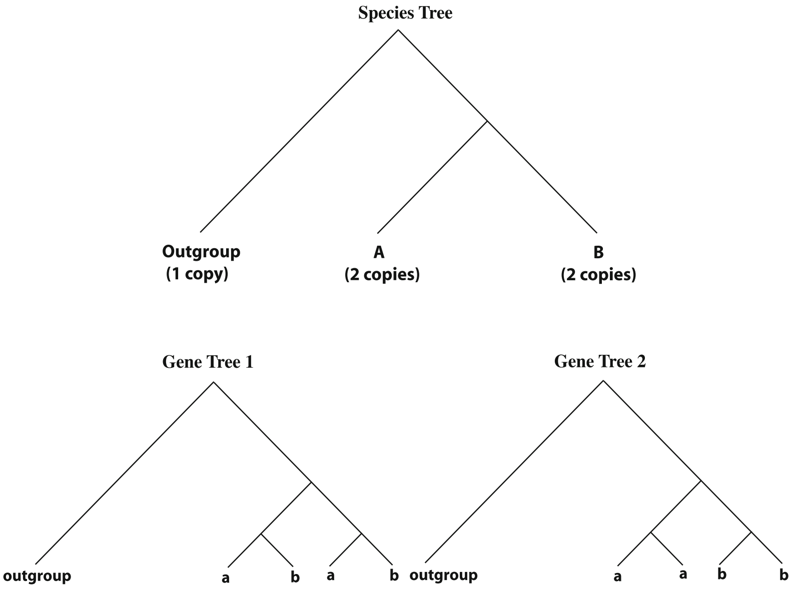


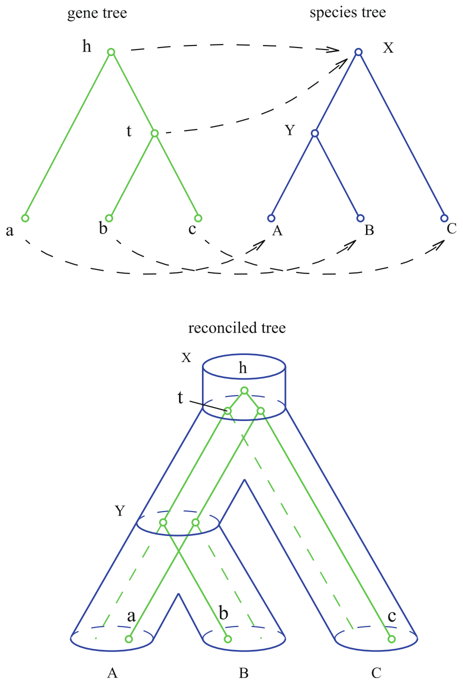
Fig. 5.1 Gene copy number data and corresponding gene trees. The gene copy number data mapped on a species tree implies at least one gene duplication, but do not specify the location of the duplication(s). Gene tree 1 implies a single duplication preceding the most recent common ancestor of A and B. Gene tree 2 implies independent duplications in the lineages leading to A and B

hybridization, or lateral gene transfer can result in gene topologies that differ from the phylogeny of the species in which the genes evolve (e.g., Maddison 1997). The general challenge of gene tree reconciliation is to find the evolutionary scenario that best explains the gene tree topologies. For the case of WGD, we might ask if a collection of gene trees is consistent with a WGD event, or at least a large number of duplications, at a particular point in the species phylogeny. While this approach seeks to directly map the location of gene duplications, and thus provide more direct evidence of WGD than simply examining changes in chromosome or gene copy numbers, in practice it often is complicated by the number of different scenarios that can cause gene tree incongruence and the inherent difficulty of accurately inferring a gene tree from only a gene sequence alignment (for more detailed, general reviews see Eulenstein et al. 2010; Doyon et al. 2011b).

5.3.1 Parsimony Approaches

Much of the initial work in gene tree reconciliation was based on optimizing a parsimony criterion; that is, finding a mapping of a gene tree topology onto the species tree that implies the fewest evolutionary events. The gene duplication model was first

Fig. 5.2 An example of lca-mapping of gene duplications. The parent and child nodes in the gene tree (h and t) map to the same node (X) in the species tree. This implies a single duplication occurring prior to or at node X in the species tree



introduced by Goodman et al. (1979; also see Page 1994 and Guigó et al. 1996) to find the minimum number of gene duplications needed to explain the incongruence between a gene tree and species tree. To do this, a gene tree can be embedded into a species tree through least common ancestor mapping (lca-mapping), which maps every node in the gene tree (tips and internal nodes) to the most recent node in the species tree that could have contained the gene node (Fig. 5.2). A duplication occurred if a parent and child node in the gene tree share the same lca-mapping in the species tree. The lca-mapping identifies the most recent possible location of the gene duplication in the species tree, but this often is not the only possible location of the gene duplication. In many cases, the duplication could have preceded the lca-mapping, although the earlier placement of the duplication may imply additional gene losses. However, as in the gene copy number analyses, gene losses are often difficult to distinguish from incomplete sampling.

This simple gene reconciliation can provide a direct estimate of the phylogenetic location of gene duplication events with a precision that is impossible with chromosome or gene copy number data. Also, it does not even require the presence of duplicated, paralogs genes; incongruence between a single-copy gene tree topology and the species phylogeny may be evidence of a hidden history of gene duplication and loss. On the other hand, gene duplications and losses are not the only explanation of gene tree incongruence. For example, incomplete lineage sorting or reticulation also can confound gene tree topologies. In this case, the gene duplication model may mistakenly imply a large number of duplications preceding rapid cladogenesis in the species tree, where we might expect high levels of incomplete lineage sorting.

Perhaps the most difficult problem underlying the gene reconciliation approaches is that, in many cases, the incongruence between a gene tree and the species phylogeny may simply be due to error. The gene tree model will interpret any topological error as evidence of duplications. Consequently, this approach often implies far more duplications rather than biologically plausible (e.g., Rasmussen and Kellis 2011). The errors in the gene tree tend to place erroneously large numbers of duplications near the root of the species tree, which may falsely suggest large-scale duplication events at the origins of major clades (Hahn 2007; Burleigh et al. 2010). In the parsimony context, several strategies have been proposed to ameliorate the problems of gene tree error. First, poorly supported clades in the gene tree may be collapsed. Several algorithmic approaches have extended the gene duplication model to deal with reconciling nonbinary trees (Berglund-Sonnhammer et al. 2006; Chang and Eulenstein 2006; Durand et al. 2006). Also, several approaches have been developed to allow minor modifications of the gene tree topology if they reduce the number of implied duplications (e.g., Chen et al. 2000; Chaudhury et al. 2011, 2012; Gorecki and Eulenstein 2012). For example, Chaudhury et al. (2012) introduced an algorithm that, given a gene tree and a species tree, finds a gene topology in a subtree pruning and regrafting (SPR) neighborhood of the original gene tree that minimizes the number of implied duplications. These local rearrangements can massively reduce the number of estimated gene duplication events.

In spite of the many issues related to gene tree reconciliation, simple and informal gene tree reconciliations have been effective at helping to identify the phylogenetic location of WGDs in plants. These approaches are usually limited to small gene trees with paralogs, that is, gene trees in which at least one duplication must have occurred. In a simple three-taxon approach, a gene tree is constructed with a pair of paralogs genes from a test taxon, and the best homologs from a sister taxon and from an outgroup taxon (e.g., Bowers et al. 2003). If paralogs from the test taxon form a clade, they diverged after the common ancestor with the sister taxon; if they do not, they diverged before the most recent common ancestor. This three-taxon phylogenetic approach provides only a limited phylogenetic context for the duplications, but it has been used to determine the timing of WGDs in *Arabidopsis* relative to its divergence from pines, rice, and other eudicots (Bowers et al. 2003) and rice relative to its divergence from pines, *Arabidopsis*, and other monocots (Vandepoele et al. 2003; Chapman et al. 2004). More recently, Jiao et al. (2011) counted the gene trees that were consistent with different scenarios of WGD to infer WGDs at the root of angiosperms and seed plants.

5.3.2 Parsimony Methods to Identify WGDs

The gene duplication problem described above treats each duplication independently. Although it may identify places in the species tree with high numbers of gene duplications, it does not attempt to find large-scale duplication events. Several proposed approaches attempt to identify the minimum number of gene duplication events, where an event may include duplications of many or all genes, rather than simply the number of duplications. One indirect approach is to examine all the possible locations of each gene duplication and find a mapping that minimizes the number of locations (nodes) on the species tree where gene duplications occur (Guigó et al. 1996; Page and Cotton 2002; Burleigh et al. 2009; Luo et al. 2009). This approach does not directly infer WGDs; but ideally, it can identify places in the species tree that are possible locations of clusters of many duplications. With a limited number of gene trees, this approach appears to be effective at identifying some locations of ancient WGDs in plants (Burleigh et al. 2009). Unfortunately, with a large number of gene trees, it is likely that all possible duplication mappings will require duplication events at every node in the species tree. In this case, every possible mapping of gene duplications will be equally optimal, and this approach will be uninformative.

Another approach seeks to find a gene duplication mapping that implies the fewest gene duplication episodes (Guigó et al. 1996; Page and Cotton 2002; Bansal and Eulenstein 2008; Luo et al. 2009). Given a single gene tree and species tree, any set of gene duplications, from the same or different gene trees, that occur on the same node in a species tree can be explained by a single gene duplication episode (or event) as long as none of the gene duplications in the set have an ancestor–descendant relationship with each other (Fig. 5.3). This approach appears to help identify WGDs, which should be very large episodes, but the largest episode is simply the largest episode found on any single gene tree (Page and Cotton 2002; Burleigh et al. 2010). In practice, the mapping that minimizes the number of episodes is determined by only the largest gene trees. Furthermore, randomizing the leaf labels (taxon names) on the gene trees can result in gene tree mappings that imply fewer episodes (Burleigh et al. 2010). Thus, although the notion of finding gene tree mappings that are consistent with large-scale duplications is desirable, it is not clear that this problem has been properly formulated.

5.3.3 Likelihood-Based Approaches

If the gene trees are accurate, the parsimony criterion for mapping gene duplications appears to perform well when the rates of duplication and loss are low (Åkerborg et al. 2009; Doyon et al. 2009). However, these approaches do not consider branch lengths in the gene or species trees, and they have a limited ability to allow multiple duplications and losses on a single branch. Perhaps more important, it is difficult to incorporate the parsimony criterion into a rigorous statistical framework to examine evolutionary hypotheses associated with gene duplication. Numerous likelihood-based models of gene duplication and loss for reconciling gene trees and species trees have been proposed (e.g., Arvestad et al. 2003, 2004, 2009; Åkerborg et al. 2009; Doyon et al. 2009, 2011a; Rasmussen and

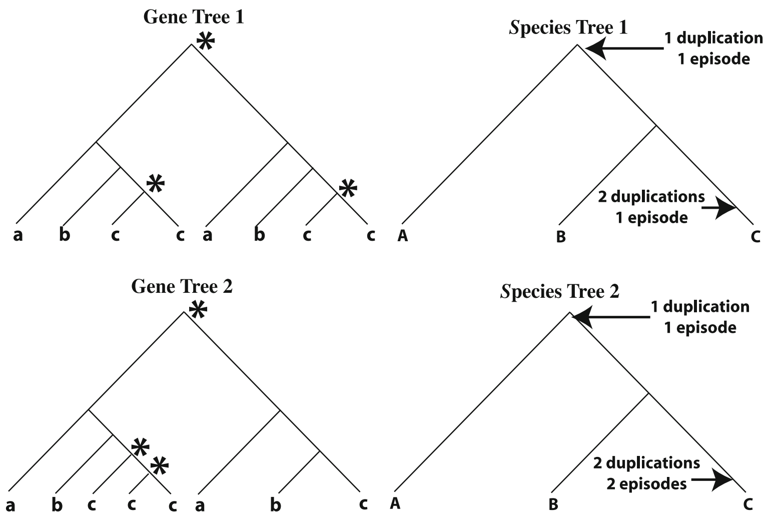


Fig. 5.3 Examples of gene duplication episodes. Duplications in the gene trees are noted with a * followed by their location in the species tree. Both gene trees have three duplications, one at the root node and two at species node C. In gene tree 1, the two duplications at species node C can be explained by a single duplication episode (since there is not a parent–child relationship between the duplication nodes). However, in gene tree 2, the two duplication episodes must have occurred at species node C since one duplication preceded the other

Kellis 2011; Gorecki et al. 2011). Most modeling approaches are based on using birth–death processes to model duplications and losses of genes as they evolve within a species tree (e.g., Arvestad et al. 2003, 2004, 2009; Åkerborg et al. 2009; Rasmussen and Kellis 2011). Also, Doyon et al. (2009, 2011a) introduced a reconciliation algorithm that calculates the likelihood of possible reconciliations based on a constant duplication and loss rate model, and this appears to produce similar reconciliations as a parsimony approach. Gorecki et al. (2011) used a Poisson model to identify the most likely distribution of reconciliations across branches in the tree based on the species branch lengths and the gene tree topologies.

Most of the likelihood model-based approaches require as input a species tree with branch lengths. The method of Gorecki et al. (2011) also requires gene tree topologies. Thus, it may be susceptible to the same problems with gene tree error as parsimony approaches. Several approaches, however, take a gene sequence alignment as input and use a Markov chain Monte Carlo (MCMC) approach to simultaneously obtain the posterior distributions of gene tree topologies and gene duplication and loss mappings (e.g., Arvestad et al. 2004, 2009; Åkerborg et al. 2009; Rasmussen and Kellis 2011). This approach provides an elegant, although computationally difficult, way to incorporate gene tree uncertainty into the gene tree reconciliation. Rasmussen and Kellis (2011) demonstrated that this approach can produce more accurate gene trees and consequently greatly reduce the number of implied duplication events compared to a parsimony approach.

The computational complexity of these likelihood-based approaches raises concerns that they may have difficulty exploiting the magnitude of new genomic data. Yet, in many ways, they still greatly simplify the complexity of genome evolution. Rasmussen and Kellis (2012) have described the first models of both duplication and coalescence, and other modeling approaches estimate the effects of hybridization and coalescence, but not duplication and loss (e.g., Meng and Kubatko 2009; Gerard et al. 2011). These represent important steps in simultaneously accounting for the many processes that affect gene tree topologies. Still, all of the modeling approaches for duplications and losses assume that the genes and all gene duplications are independent. Thus, they may detect branches in the species tree with high rates of duplication or loss, but they do not directly assess the likelihood of WGDs. The development of such models that allow for simultaneous duplications across many genes can allow for rigorous statistical tests of the placement of WGD events.

5.4 Other Genomic Data

From availability of large-scale genomic data, evidence of cryptic ancient WGDs often comes from either identifying large, syntenic (duplicated) blocks within a single genome, or by looking at the age distribution of duplicated genes within a chromosome (see Van de Peer 2004). Since these approaches use only data from a single species and are focused more on identifying ancient WGDs than placing the WGDs in an evolutionary context, I will not cover them in detail. The presence of duplicated chromosomal segments may provide direct, unambiguous evidence of WGDs that may be difficult to obtain from simply gene copy numbers or gene trees (Vision et al. 2000). However, in practice, rapid gene losses and rearrangements after polyploidy can make it extremely difficult to detect such duplications, and different methods of detecting duplicated blocks and using different criteria for defining a syntenic block can greatly affect interpretations of the history of large-scale duplications (see Durand and Hoberman 2006). Although simply examining the genome of a single species cannot reveal the phylogenetic context of a WGD, the dates of the ancient divergences can be estimated based on the molecular divergence of paralogs. It may be possible to map the evolution of large duplicated segments on a tree, but in plants, this may require extending the taxonomic sampling of species with adequate genomic mapping data. Perhaps the greater contribution of these duplicated regions is that they can define sets of paralogs that originated from WGDs, and this information can be used to validate mappings of duplications from gene copy number or gene reconciliation analyses.

The rapidly increasing abundance of large-scale transcriptome data sets for plants provides an opportunity to define WGDs based on the age distribution of duplicated genes (see Cui et al. 2006). The methods first can define pairs of most recent gene duplicates using methods such as an all-by-all BLAST. If gene duplication and loss occur at a constant rate, the frequency of duplicated genes in a genome will decrease exponentially with time. In contrast, a large-scale duplication event, like a WGD, should result in an overrepresentation of duplicated gene pairs at the time corresponding to the large-scale duplication event. In practice, in a plot of the age distribution, usually represented by synonymous substitution distance, of duplicated genes, peaks in the age distribution curves or evidence of multiple distributions, may indicate WGDs. Again, it is difficult to precisely place a WGD just from the pairwise divergence of duplicated sequences, but with data available from many taxa, comparisons of these age plots from related species can be informative. In some cases, analyses of the age distribution plots have failed to detect known WGDs (e.g., Blanc and Wolfe 2004; Paterson et al. 2004). However, unlike gene tree reconciliation methods, they will not be misled by incomplete lineage sorting or gene tree error.

5.5 Conclusions

New genomic sequence data and advances in phylogenetic methods presents unprecedented opportunities to place WGDs within the plant tree of life. Still, there is much work to do. Although numerous data sources and methods may provide evidence of WGDs, there exist few statistical tests of WGD. A rigorous statistical framework still must be developed to examine hypotheses about the locations of WGDs. Also, examinations of the phylogenetic placement of WGDs often are based on available data; data sets are rarely generated solely for the purpose of placing the location of WGDs. Thus, there has been little discussion about the optimal methods or optimal data sets for mapping WGDs. Indeed, this is a complex issue. For example, gene trees may allow direct observations of the patterns of gene duplication and loss, but they also are susceptible to many errors and biases that may not be problems with simpler data, such as gene copy number. Ideally, learning more about the evolutionary context and implications of WGDs in plants (e.g., their effect on diversification rates and their relationship to phenotypes such as life history or mating system) will also help to identify and place WGDs in plants.

Acknowledgments This chapter was developed and written in parts with support from the Gene Tree Reconciliation Working Group at NIMBioS through NSF award EF-0832858, with additional support from the University of Tennessee. Discussions with members of the working group including Cecile Ané, Oliver Eulenstein, Pawel Gorecki, and Brian O’Meara were helpful for this manuscript.

References

Åkerborg Ö, Sennlad B, Arvestad L, Lagergren J (2009) Simultaneous bayesian gene tree reconstruction and reconciliation analysis. Proc Natl Acad Sci USA 106:5714–5719

Ames RM, Money D, Ghatge VP, Whelan S, Lovell SC (2012) Determining the evolutionary history of gene families. Bioinformatics (In press)

Ané C (2008) Analysis of comparative data with hierarchical autocorrelation. Ann Appl Stat 2:107–1102

Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2003) Bayesian gene/species tree reconciliation and orthology analysis using MCMC. Bioinformatics 19:i7–i15

Arvestad L, Berglund A-C, Lagergren J, Sennblad B (2004) Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. RECOMB 2004:326–335

Arvestad L, Lagergren J, Sennblad B (2009) The gene evolution model and computing its associated probabilities. J ACM 56:7

Bansal MS, Eulenstein O (2008) The multiple gene duplication problem revisited. Bioinformatics 24:i132–i138

Berglund-Sonnhammer A-C, Steffansson P, Betts MJ, Liberles DA (2006) Optimal gene-trees from sequences and species trees using a soft interpretation of parsimony. J Mol Evol 63:240–250

Blanc G, Wolfe KH (2004) Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1093–1101

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

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

Buggs RJA, 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

Buggs RJA, Chamala S, Wu W, Tate JA, Schnable PS, Soltis DE, Soltis PS, Barbazuk WB (2012) Rapid, repeated, and clustered loss of duplicated genes in allopolyploid plant populations of independent origin. Curr Biol 22:1–5

Burleigh JG, Bansal MS, Wehe A, Eulenstein O (2009) Locating large-scale gene duplication events through reconciled trees: implications for identifying ancient polyploidy in plants. J Comput Biol 16:1071–1083

Burleigh JG, Bansal M, Eulenstein O, Vision TJ (2010) Inferring species trees from gene duplication episodes. Proc BCB 2010:198–203

Chang W-C, Eulenstein O (2006) Reconciling gene trees with apparent polytomies. COCOON 2006. LNCS 4112:235–244

Chapman BA, Bowers JE, Schulze SR, Paterson AH (2004) A comparative phylogenetic approach for dating whole genome duplication events. Bioinformatics 20:180–185

Chaudhary R, Burleigh JG, Eulenstein O (2011) Algorithms for rapid error correction for the gene duplication problem (ISBRA) 2011. LNCS 6674:184−196

Chaudhary R, Burleigh JG, Eulenstein O (2012) Efficient error correction algorithms for gene tree reconciliation based on duplication, duplication and loss, and deep coalescence. BMC Bioinformatics 13:s11

Chen K, Durand D, Farach-Colton M (2000) Notung: a program for dating gene duplications and optimizing gene family trees. J Comput Biol 7:429–447

Chester M, Gallagher JP, Symonds VV, Cruz da Silva AV, Mavrodiev EV, Leitch AR, Soltis PS, Soltis DE (2012) Extensive chromosomal variation in a recently formed natural allopolyploid species, *Tragapogon miscellus* (Asteraceae). Proc Nat Acad Sci USA 109:1176–1181

Csurös M (2010) Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26:1910–1912

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, de Pamphilis CW (2006) Widespread genome duplications throughout the history of flowering plants. Genome Res 16:738–749

Cusimano N, Sousa A, Renner SS (2012) Maximum likelihood inference implies a high, not a low, ancestral haploid chromosome number in araceae, with a critique of the bias introduced by ‘x’. Ann Bot 109:681−692

De Bie T, Cristianini N, Demuth JD, Hahn MW (2006) CAFÉ: a computational tool for the study of gene family evolution. Bioinformatics 22:1269–1271

DeBolt S (2010) Copy number variation shapes genome diversity in A*rabidopsis* over immediate family generational scales. Genome Biol Evol 2:441–453

Digby L (1912) The cytology of *Primula kewensis* and of other related *Primula* hybrids. Ann Bot 26:357–388

Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF (2008) Evolutionary genetics of genome merger and doubling in plants. Annu Rev Genet 42:443–461

Doyon J-P, Chauve C, Hamel S (2009) Space of gene/species tree reconciliations and parsimonious models. J Comput Biol 16:1399–1418

Doyon J-P, Hamel S, Chauve C (2011a) An efficient method for exploring the space of gene tree/ species tree reconciliations in a probabilistic framework. IEEE/ACM Trans. Comput Biol Bioinform 99: (In press)

Doyon J-P, Ranwez V, Daubin V, Berry V (2011b) Models, algorithms and programs for phylogeny reconciliation. Briefings Bioinform 12:392–400

Durand D, Halldórsson B, Vernot B (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. J Comput Biol 13:320–335

Durand D, Hoberman R (2006) Diagnosing duplications—can it be done? Trends Genet 22:156–164

Ehrendorfer F, Krendl F, Habeler E, Sauer W (1968) Chromosome numbers and evolution in primitive angiosperms. Taxon 17:337–468

Eulenstein O, Huzurbazar S, Liberles DA (2010) Reconciling phylogenetic trees. In: Dittmar K, Liberles D (eds) Evolution after gene duplication. Wiley, Hoboken, pp 185–206

Gaut BS (2001) Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Res 11:55–66

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

Gerard D, Gibbs HL, Kubatko L (2011) Estimating hybridization in the presence of coalescence using phylogenetic intraspecific sampling. BMC Evol Biol 11:291

Goldblatt P (1980) Polyploidy in angiosperms: monocotyledons. In: Lewis WH (ed) Polyploidy: biological relevance. Plenum Press, New York, pp 219–239

Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed by globin sequences. Syst Zool 28:132–163

Gorecki P, Eulenstein O (2012) Simultaneous error correction and rooting for gene tree reconciliation and the gene duplication problem. BMC Bioinformatics (In press)

Gorecki P, Eulenstein O, Burleigh JG (2011) Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. BMC Bioinform 12:S15

Grant V (1963) The origin of adaptations. Columbia University Press, New York

Grant V (1982) Periodicities in the chromosome numbers of the angiosperms. Bot Gaz 143:379–389

Guggisberg A, Mansion G, Kelso S, Conti E (2006) Evolution of biogeographic patterns, ploidy levels, and breeding systems in a diploid-polyploid species complex in primula. New Phytol 171:617–632

Guigó R, Muchnik I, Smith TF (1996) Reconstruction of ancient molecular phylogeny. Mol Phylogenet Evol 6:189–213

Hahn MW (2007) Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. Genome Biol 8:R141

Hahn MW, De Bie T, Stajich JE, Nguyen C, Cristianni N (2005) Estimating the tempo and mode of gene family evolution from comparative data. Genome Res 15:1153–1160

Hipp AL, Rothrock PE, Reznicek AA, Berry PE (2007) Chromosome number changes associated with speciation in sedges: a phylogenetic study in *Carex* section *Ovales* (Cyperaceae) using AFLP data. Aliso 23:193–203

Iwasaki W, Takagi T (2007) Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics 23:i230–i239

Jiao Y, Wickett NJ, Ayampalayam 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, de Pamphilis CW (2011) Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–102

Kunin V, Ouzounis CA (2003) GeneTRACE-reconstruction of gene content of ancestral species. Bioinformatics 19:1412–1416

Levin DA (2002) The role of chromosomal change in plant evolution. Oxford University Press, New York

Librado P, Vieira FG, Rozas J (2012) BadiRate: estimating family turnover rates by likelihood-based methods. Bioinformatics 28:279–281

Lim KY, Soltis DE, Soltis PS, Tate J, Matyasek R, Srubarova H, Kovarik A, Pires JC, Xiong Z, Leitch AR (2008) Rapid chromosome evolution in recently formed polyploids in *Tragapogon* (Asteraceae). PLoS ONE 3:e3353

Liu L, Yu L, Kalavacharla V, Liu Z (2011) A bayesian model for gene family evolution. BMC Bioinform 12:426

Luo CW, Chen MC, Chen YC, Yang RWL, Liu HF, Chao KM (2009) Linear-time algorithms for the multiple gene duplication problems. IEEE/ACM Trans Comput Biol Bioinform 99:5555

Maddison WP (1997) Gene trees in species trees. Syst Biol 46:523–536

Masterson J (1994) Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science 264:421–424

Mayrose I, Barker MS, Otto SP (2010) Probabilistic models of chromosome evolution and the inference of polyploidy. Syst Biol 59:132–144

Mayrose I, Zhan SH, Rothfels CJ, Magnus-Ford K, Barker MS, Rieseberg LH, Otto SP (2011) Recently formed polyploidy plants diversify at lower rates. Science 333:1257

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75:35–45

Meyers LA, Levin DA (2006) On the abundance of polyploids in flowering plants. Evolution 60:1198–1206

Mirkin BG, Fenner TI, Galperin MY, Koonin EV (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3:2

Mishima M, Ohmido N, Fukui K, Yahara T (2002) Trends in site-number change of rDNA loci during polyploidy evolution in *Sanguisorba* (Rosaceae). Chromosoma 110:550–558

Page RDM (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. Syst Biol 43:58–77

Page RDM, Cotton JA (2002) Vertebrate phylogenomics: reconciled trees and gene duplication. Pac Symp Biocomput, 536–547

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

Rasmussen MD, Kellis M (2011) A Bayesian approach for fast and accurate gene tree reconstruction. Mol Biol Evol 28:273–290

Rasmussen MD, Kellis M (2012) Unified modeling of gene duplication, loss and coalescence using a locus tree. Genome Res 22:755−765

Raven PH (1975) The bases of angiosperm phylogeny: cytology. Ann Mo Bot Gard 62:724–764

Schluter D, Price T, Mooers AØ, Ludwig D (1997) Likelihood of ancestor states in adaptive radiation. Evolution 41:1239–1251

Schultheis LM (2001) Systematics of *Downingia* (Campanulaceae) based on molecular sequence data: implications for floral and chromosome evolution. Syst Bot 26:603–621

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

Soltis PS, Soltis DE (2000) The role of genetic and genomic attributes in the success of polyploids. Proc Natl Acad Sci 97:7051–7057

Snel B, Bork P, Huynen MA (2002) Genomes in flux: the evolution of archael and proteobacterial gene content. Genome Res 12:17–25

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

Stace HM, Chapman AR, Lemson KL, Powell JM (1997) Cytoevolution, phylogeny, and taxonomy in Epacridaceae. Ann Bot 79:283–290

Stebbins GL (1938) Cytological characteristics associated with the different growth habits in the dicotyledons. Am J Bot 25:189–198

Stebbins GL (1950) Variation and evolution in plants. Columbia University Press, New York

Stebbins GL (1971) Chromosomal evolution in higher plants. Addison-Wesley, London

Tate JA, Ni Z, Scheen A-C, Koh J, Gilbert CA, Lefkowitz D, Chen ZJ, Soltis PS, Soltis DE (2006) Evolution and expression of homeologous loci in *Tragopogon miscellus* (Asteraceae), a recent and reciprocally formed allopolyploid. Genetics 173:1599–1611

Vamosi JC, Dickinson TA (2006) Polyploidy and diversification: a phylogenetic investigation in Rosaceae. Int J Plant Sci 167:349–358

Vandepoele K, Simillion C, Vande Peer Y (2003) Evidence that rice and other cereals are ancient aneuploids. Plant Cell 15:2192–2202

Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. Nat Rev Genet 5:752–763

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

Walker JW (1972) Chromosome numbers, phylogeny, phytogeography of the Annonaceae and their bearing on the (original) basic chromosome number of angiosperms. Taxon 21:57–65

Winge Ö (1917) The chromosomes. Their numbers and general importance. Comptes Rendus des Travaux Laboratoire Carlsberg 13:131–275

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

Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH (2009) The frequency of polyploidy speciation in vascular plants. Proc Natl Acad Sci USA 106:13875–13879

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, Liu T-F, Jiang S, Ramachandran S, Liu C-M, Jing H-C (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum *(Sorghum bicolor*). Genome Biol 12:R114