Evaluating and Characterizing Ancient Whole-Genome Duplications in Plants with Gene Count Data

George P. Tiley[1,*], Cécile Ané[2,3], and J. Gordon Burleigh[1]

[1]Department of Biology, University of Florida

[2]Department of Statistics, University of Wisconsin-Madison

[3]Department of Botany, University of Wisconsin-Madison

*Corresponding author: E-mail: gtiley@ufl.edu.

Accepted: March 11, 2016

Abstract

Whole-genome duplications (WGDs) have helped shape the genomes of land plants, and recent evidence suggests that the genomes of all angiosperms have experienced at least two ancient WGDs. In plants, WGDs often are followed by rapid fractionation, in which many homeologous gene copies are lost. Thus, it can be extremely difficult to identify, let alone characterize, ancient WGDs. In this study, we use a new maximum likelihood estimator to test for evidence of ancient WGDs in land plants and estimate the fraction of new genes copies that are retained following a WGD using gene count data, the number of gene copies in gene families. We identified evidence of many putative ancient WGDs in land plants and found that the genome fractionation rates vary tremendously among ancient WGDs. Analyses of WGDs within Brassicales also indicate that background gene duplication and loss rates vary across land plants, and different gene families have different probabilities of being retained following a WGD. Although our analyses are largely robust to errors in duplication and loss rates and the choice of priors, simulations indicate that this method can have trouble detecting multiple WGDs that occur on the same branch, especially when the gene retention rates for ancient WGDs are very low. They also suggest that we should carefully evaluate evidence for some ancient plant WGD hypotheses.

Key words: whole-genome duplication, gene duplication and loss, gene family evolution, paleopolyploidy, gene count, gene retention.

Introduction

Whole-genome duplications (WGDs) play an important role in shaping the genomes of plants (e.g., Adams and Wendel 2005). The availability of large-scale genomic data and fully sequenced genomes has revealed much evidence for ancient WGDs or paleopolyploidy (e.g., Vision et al. 2000; Cui et al. 2006; Jaillon et al. 2007; Jiao et al. 2011; Vanneste et al. 2014). In fact, recent evidence suggests at least two WGDs preceded the diversification of angiosperms (Jiao et al. 2011; *Amborella* Genome Project 2013; Li et al. 2015). WGDs in plants often are followed by rapid fractionation, in which many homeologous gene copies are lost (Otto 2007; Mandáková et al. 2010; Schnable et al. 2012), and this diminishes evidence of WGDs over time. Consequently, although WGDs appear to be pervasive throughout the evolutionary history of plants, it can be extremely difficult to detect, let alone characterize, these ancient events (e.g., Burleigh 2012).

Perhaps the most direct evidence for ancient WGDs is the presence of large syntenic regions within a genome (e.g., Jaillon et al. 2004; Kellis et al. 2004). However, few studies have performed statistically rigorous tests of WGD hypotheses based on syntenic data, and many measures of synteny are presented without estimates of uncertainty. Although recent models of gene family evolution use syntenic data and account for many complexities of WGD, these are used for probabilistic orthology prediction and not explicitly testing WGDs (Conant and Wolfe 2008; Conant 2014). Additionally, reoccurring WGDs throughout plant evolution can make the interpretation of syntenic data difficult, especially without well-assembled genomes that can be used to detect synteny between species.

In the absence of well-annotated genomes assembled to the chromosome level, ancient WGDs often are inferred from the distribution of the rate of synonymous substitution per synonymous site (d*S*) among duplicate genes in a genome (e.g., Lynch and Conery 2000; Raes et al. 2003). It is generally assumed that gene duplication and loss follows a steady-state birth–death process, with constant rates of duplication (birth)

or loss (death) per gene family (Lynch and Conery 2003). WGDs violate this assumption, and consequently, WGDs produce peaks in cumulative distributions of pairwise d*S* between paralogs within a genome (e.g., Gu et al. 2002; McLysaght et al. 2002; Jaillon et al. 2004; Vandepoele et al. 2004; Maere et al. 2005). In the case of an ancient WGD, the steady-state birth–death process alone cannot describe the distribution of d*S* between paralogs. However, the signal of a WGD in the distribution of d*S* may degrade through time, and it can be extremely difficult to identify ancient WGDs based on d*S* distributions (e.g., Blanc and Wolfe 2004a; Paterson et al. 2004). Furthermore, multiple substitutions at the same site can create peaks in d*S* plots that do not correspond to WGDs (Vanneste et al. 2013).

Recently, Rabier et al. (2014) described a new approach to identify WGDs (or WGTs [whole-genome triplications]) on a phylogeny based on gene count data, the number of gene copies in various gene families across a group of taxa. Hahn et al. (2005) originally developed a maximum likelihood approach to estimate gene birth (i.e., duplication) and death (i.e., loss) rates on a phylogeny with gene count data. Rabier et al. (2014) extended this approach to estimate background rates of gene duplication (λ) and loss (μ) throughout the tree and account for the probability of a WGD when reconstructing ancestral gene copy numbers along the nodes of a phylogeny. When the user defines WGDs (or WGTs) within the tree, the model also assumes a proportion of the extra duplicated genes are lost immediately following the WGD event. The fraction of extra genes that survive from a WGD is the retention rate (q), and the model estimates independent gene retention rates for each WGD in the tree. This approach is appealing because the user can explicitly test for WGDs along specific branches of a phylogeny with a likelihood ratio test by comparing models with and without WGDs. The user also can estimate the timing of the WGD along a branch; the likelihood is maximized when the WGD is placed in its optimal position along an edge in the species phylogeny.

In this study, we evaluate the evidence for numerous ancient WGDs across land plants and estimate gene retention rates following WGDs using gene copy number data from fully sequenced genomes. We explore the effects of estimates of gene duplication rates (λ) and gene loss rates (μ) on estimates of WGD retention rates (q) and use simulations and empirical tests to assess our ability to detect WGD events across sequenced plant genomes.

Materials and Methods

Gene Family Data

We obtained gene counts from 30,023 orthogroup clusters circumscribed by the *Amborella* Genome Project (2013) using OrthoMCL (Li et al. 2003). We first filtered the gene count data to remove any families that did not span the root of the land plant phylogeny to eliminate any gene families that arose de novo within the land plants (fig. 1). This land plant data set includes gene families that have at least one copy in *Physcomitrella patens* and at least one copy in another taxon. Failing to condition gene count data with at least one copy in each clade spanning the root of the phylogeny can lead to biased estimates of λ and μ (Rabier et al. 2014). We also created gene count data sets that were filtered to include at least one copy in *Amborella trichopoda* and at least one copy in a monocot lineage (the monocot data set), at least one copy in *A. trichopoda* and another in a eudicot lineage (the eudicot data set), and one in *Theobroma cacao* and another in a Brassicales taxon (the Brassicales data set; fig. 1). Likelihood calculations based on probabilities of gene count data can be extremely memory intensive when there are no limits to ancestral gene family sizes. Therefore, we removed gene families with $\geq$100 copies in any taxon from all gene count data sets, and we set the ancestral gene family size to a maximum of 100. The filtering process resulted in data from 7,564, 10,795, 11,249, and 12,957 gene families in the land plant, monocot, eudicot, and Brassicales gene count data sets, respectively.

Phylogenetic Tree

We used a species tree with a topology that corresponds to our current understanding of land plant relationships (fig. 1; e.g., Soltis et al. 2011; Ruhfel et al. 2014). The relationships of the taxa used here are generally well supported, although there is some disagreement about the position of *Populus* (see Sun et al. 2015). We obtained estimates of the ultrametric branch lengths for most branches in the angiosperms based on the dating analysis of Bell et al. (2010), who used a BEAST analysis (Drummond and Rambaut 2007) based on 36 minimum fossil age constraints, which were treated as exponential distributions. Divergence times for *Physcomitrella* and the other land plants, *Selaginella* and the angiosperms, and the two *Solanum* species were obtained from TimeTree (http://www.timetree.org [last accessed July 21, 2013]; Hedges et al. 2006). The species tree with branch lengths is available as supplementary data, Supplementary Material online. We used ultrametric branch lengths in Myr, since we assume gene duplication and loss is a function of calendar time as opposed to evolutionary time measured, for instance, as a number of substitutions per sites from selected genes.

Timing of Hypothesized WGD Events

We identified hypotheses for ancient WGD events along the land plant phylogeny from primary literature. Support for these hypotheses comes from various lines of evidence, including synteny, d*S* plots, and mapping gene duplication events onto a phylogeny.

There is much evidence in support of multiple WGDs in the evolutionary history of *Arabidopsis thaliana* (e.g., *Arabidopsis* Genome Initiative 2000; Blanc et al. 2000; Vision et al. 2000;

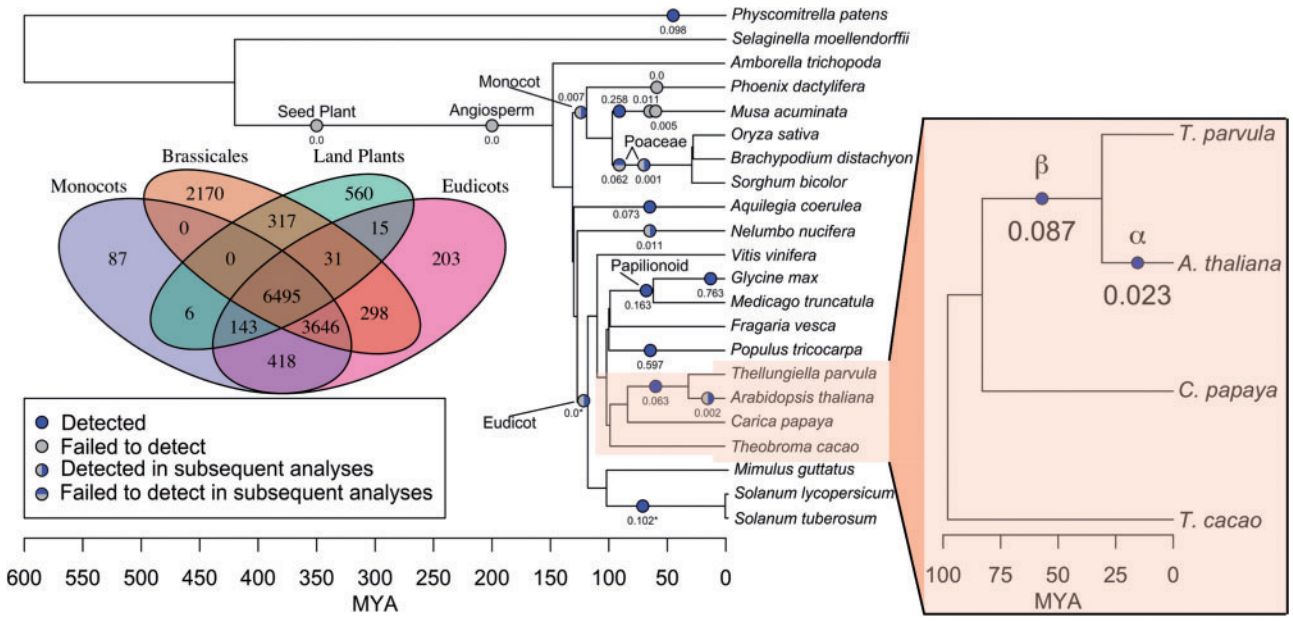

Fig. 1.—Ultrametric land plant phylogeny with putative ancient WGD or WGT events plotted as circles on branches. Numbers displayed next to events are retention rates (q) optimized using a prior geometric mean of 1.5. Events associated with significant LRTs for analyses with land plant data are in blue and nonsignificant events are colored gray. Some events are detected in subsequent analyses but not the land plant analyses, while some events are not consistently detected across all analyses. An "*" next to a retention rate indicates a WGT rather than a WGD. The Venn diagram shows how many gene families were used for each data set and the overlap between data sets. The data sets depicted in the Venn diagram refer to the different data filtering strategies. A four-taxon tree formed by *Theobroma cacao* and Brassicales is shown in the zoomed-in portion with the *Arabidopsis* α and β WGD events. 12,957 gene families span the root of the four-taxon tree (the Brassicales data), and optimizing model parameters causes a significant LRT for both *Arabidopsis* α and β as well, as higher retention rates.

Raes et al. 2003). The most recent WGD in the *Ar. thaliana* lineage (*Arabidopsis* α) is thought to be shared with most members of the Brassicaceae (The *Brassica rapa* Genome Sequencing Project Consortium 2011); however, based on the *Carica papaya* genome, it is not shared across the Brassicales (Ming et al. 2008). An older WGD event in the *Ar. thaliana* lineage (*Arabidopsis* β) also is not shared by *C. papaya* (Ming et al. 2008; Argout et al. 2011) but is shared by all sequenced Brassicaceae (The *Brassica rapa* Genome Sequencing Project Consortium 2011; Dassanayake et al. 2011). Bowers et al. (2003) estimated *Arabidopsis* α at 14.5–20.4 Ma based on the shared syntenic regions with other partially sequenced angiosperm genomes. More recently, analyses of syntenic regions shared between *Ar. thaliana* and *Thellungiella parvula* revealed that the *Arabidopsis* α WGD occurred prior to the divergence of *Arabidopsis* and *Thellungiella* (Dassanayake et al. 2011), and Vanneste et al. (2014) dated the *Arabidopsis* α WGD at approximately 48 Ma. We originally tested the *Arabidopsis* α and β on the branch prior to the divergence of *Ar. thaliana* and *Thel. parvula*, with *Arabidopsis* α placed at 48 Ma and *Arabidopsis* β placed at 82.999999 Ma. However, we could not detect *Arabidopsis* β under any conditions when *Arabidopsis* α and β were on the same branch (supplementary tables S5 and S10, Supplementary Material online). Therefore, we placed *Arabidopsis* α on the tip leading to *Ar. thaliana* at 15.5 Ma (fig. 1). The timing of *Arabidopsis* β is uncertain, so we placed it at 57 Ma, the midpoint of the branch leading to Brassicaceae (fig. 1). Placing the *Arabidopsis* α and β WGDs on two separate branches allowed us to then detect both WGDs on the four-taxon tree of *T. cacao* and Brassicales.

Populus trichocarpa has over 45,000 genes, approximately 8,000 of which are thought to be retained from a recent WGD (Tuskan et al. 2006). Comparisons with *Salix* ESTs indicate the WGD occurred before the divergence of *Populus* and *Salix*, but this also suggests *Po. trichocarpa* has a slow rate of synonymous substitution when compared to *Arabidopsis* (Tuskan et al. 2006). Estimating the age of the *Po. trichocarpa* WGD based on the distribution of d*S* between *Po. trichocarpa* paralogs with the *Ar. thaliana* mutation rate calibration (Lynch and Conery 2000) places the WGD event at 8–13 Ma, after the divergence of *Populus* and *Salix* (Tuskan et al. 2006). Therefore, we placed the putative *Po. trichocarpa* WGD at 49.5 Ma, the midpoint of the *Po. trichocarpa* branch on the species tree.

WGDs are prevalent across the Fabaceae (Cannon et al. 2015), including a lineage-specific WGD in *Glycine max* (Schmutz et al. 2010) and another WGD at the root of

Papilionoideae (Schmutz et al. 2010; Young et al. 2011). Comparison of gene pairs on syntenic segments in *G. max* and other genomes indicated that these events occurred around 13 and 59 Ma, respectively (Schmutz et al. 2010). We tested the *G. max* WGD at 13 Ma and moved the Papilionoideae WGD back to 68 Ma so it could be on the branch preceding the divergence of *G. max* and *Medicago truncatula* in our species tree.

Comparisons of syntenic gene pairs shared between *Solanum lycopersicum* and *Solanum tuberosum* indicate a WGT predating the divergence of these lineages (The Tomato Genome Consortium 2012). The age distribution of d*S* between syntenic paralogs in *S. lycopersicum* dates this event anywhere from 52 to 90 Ma. We placed the WGT at 71 Ma, on the branch shared by the *S. lycopersicum* and *S. tuberosum* lineages.

Columbine tetraploidy was proposed based on the distribution of d*S* from 178 paralogous gene pairs (Cui et al. 2006). This was one of the first studies to identify a potential ancient WGD in the Ranunculales. It is difficult to estimate an age for this putative WGD event because there are no other species with genomic data along a lineage to provide some bounds for the event. Therefore, we placed the Ranunculales WGD at 64.5 Ma, halfway along the tip leading to *Aquilegia coerulea* (the only Ranunculales taxon in the tree).

Both syntenic evidence and distributions of d*S* support a WGD specific to the *Nelumbo* lineage, that is not shared by *Vitis*, and is much more recent than the WGD common to all angiosperms (Ming et al. 2013). Dating of paralogous genes in *Nelumbo nucifera* suggests the hypothesized WGD occurred 54–76 Ma. Therefore, we placed the *Nelumbo* WGD at 65 Ma. The *Nelumbo* WGD was only tested in analyses where the number of genes at the root of the species tree for each gene family was distributed as a geometric distribution with a mean of 1.5.

Comparisons of paralogs in *Vitis vinifera* and their homologs in *Oryza sativa*, *Po. trichocarpa*, and *Ar. thaliana* suggest a paleohexiploidization event preceding the divergence of *V. vinifera* and eurosids (Jaillon et al. 2007). Observations of ratios of syntenic blocks across plant genomes also suggest a WGT before the divergence of eudicots (Lyons et al. 2008; Argout et al. 2011; *Amborella* Genome Project 2013) in addition to evidence from gene trees (Jiao et al. 2012).

Evidence suggests two WGDs (ρ and σ) may have preceded the diversification of grasses. The ρ WGD was estimated to occur around 70 Ma, and the σ WGD was estimated to have occurred around 130 Ma (Tang et al. 2010). Both of these WGDs are supported by syntenic data, and the age of each event was estimated using distributions of d*S*. However, estimating the precise timing of older WGDs, such as Poaceae σ, is difficult due to saturation of synonymous substitutions. Thus, we placed σ near 96 Ma, close to the divergence of *Musa acuminata* and Poaceae. Comparisons of the *O. sativa* and *V. vinifera* genomes by Tang et al. (2010) revealed some shared synteny, suggesting an additional WGD may have occurred in the monocot lineage before Poaceae σ. We tested this hypothesis using gene count probabilities by placing a WGD on the branch leading to the most recent common ancestor of monocots near 130 Ma.

Three WGDs in the *M. acuminata* lineage after the split between Zingiberales and Poales also have been proposed (D'Hont et al. 2012). Syntenic blocks of paralogous genes within *M. acuminata* suggest these three duplications are not shared with the other grass species in our tree, and the distribution of d*S* between paralogous gene pairs dates two duplications to approximately 65 Ma (the most recent is referred to as *M. acuminata* α and the next *M. acuminata* β), near the Cretaceous–Tertiary boundary. The third duplication (*M. acuminata* γ) is estimated to have occurred approximately 100 Ma based on the age of paralogs in 12 syntenic blocks (D'Hont et al. 2012). These syntenic blocks are not homologous to the syntenic blocks in grasses that indicate the grass ρ and σ duplications (Tang et al. 2010).

An analysis of the distributions of d*S* of *M. acuminata*, *O. sativa*, *Sorghum bicolor*, *Brachypodium distachyon, Phoenix dactylifera*, and *Ar. thaliana* proteomes based on predicted gene models suggests an additional WGD within the *Ph. dactylifera* lineage that occurred after the divergence of Zingiberales and Arecales (Al-Mssallem et al. 2013). In the *Ph. dactylifera* genome sequence, 4,215 genes out of 41,660 annotated gene models were paralogous and were arranged in 411 collinear blocks. There was a bimodal distribution *of* d*S*, but it is uncertain if this is due to a *Ph. dactylifera*-specific WGD or an older shared WGD, such as a WGD predating monocots or angiosperms (Al-Mssallem et al. 2013). Because of the uncertainty in the date of this event, we tested the hypothesized WGD at 59 Ma on the branch leading to *Ph. dactylifera*.

Phylogenetic approaches dating of gene families inferred that WGDs took place before the diversification of angiosperms and seed plants. Based on different orthogroup filtering methods, the WGD preceding angiosperms was estimated to have occurred at 192, 210, or 234 Ma, while the WGD predating seed plants was estimated to have occurred at 319, 321, or 347 Ma (Jiao et al. 2011). Comparison of *A. trichopoda* genome assembly to itself revealed 47 syntenic blocks, which contained 466 gene pairs (*Amborella* Genome Project 2013). Shared synteny with *V. vinifera* suggests that the WGD predated the divergence of *Amborella* and other angiosperms, but there was no syntenic evidence of the duplication predating seed plants (*Amborella* Genome Project 2013). The angiosperm WGD was set at 200 Ma, and the age of the seed plant WGD was set at 350 Ma (but see subsequent analyses in section "Exploring the Timing of Ancient WGD events" in which we tested different dates).

Evidence for a WGD in the *P. patens* lineage is based on a distribution of d*S* from paralogs in the *P. patens* genome (Rensing et al. 2007). The age of the *P. patens* WGD was

estimated to be 45 Ma, which is where we placed it in our tree.

WGD Retention and Loss Rates

We used the R package WGDgc (Rabier et al. 2014), run with R version 3.0.2 (R Development Core Team 2013), to test hypotheses of ancient WGDs across land plants and to estimate the rates of retaining duplicated genes following independent WGD events. The land plant, eudicot, monocot, and Brassicales data sets were run under a range of priors. Specifically, the number of genes at the root of the tree is assumed to be a geometrically distributed random variable, and the expected number of genes at the root is our prior parameter. Based on distributions of average gene family sizes (fig. 2*a*), gene duplication and loss rates (fig. 2*b*), and the likelihood scores given the data and models, the most appropriate prior was a geometric mean of 1.5. Having a prior geometric mean >1 allows for more than one gene copy from a gene family at the root of the tree, while not favoring too many copies at the root. When using a higher prior, gene birth rates go to 0 in the monocot and Brassicales data sets (fig. 2*b*). We found that estimated retention rates following WGDs are relatively robust to the choice of prior (supplementary tables S1–S4, Supplementary Material online), which is consistent with simulations in Rabier et al. (2014).

The likelihood function is calculated as $L(D|\lambda, \mu, e, q)$, where D is the set of gene family sizes $\{D_1, D_2, \ldots, D_n\}$, λ and μ are the gene duplication and loss rates, e is the set of WGD and WGT events of known placement on the tree, and q is the set of retention rates at these WGD events. Gene families are assumed to evolve independently. Therefore, the likelihood $\Pi_{i=1}^{n} L(D_i|\lambda, \mu, e, q)$is the product of likelihoods for individual gene families. The likelihood for a single gene family is $\mathbb{P}\{D_i\}/\mathbb{P}\{\text{family } i \text{ isretained}\}$.$\mathbb{P}\{D_i\}$ is calculated on a given species tree from the tips to the root using a postorder tree traversal similarly to Felsenstein's pruning algorithm (Felsenstein 1981), except that all values for the number of genes must be integrated at each node in the tree. On edges that do not include a WGD (or WGT) event, this algorithm was described by Csűrös and Miklós (2009). WGD events break edges into segments that include the background duplication and loss of genes only and segments that include WGD events only (Rabier et al. 2014). On edge segments with a single WGD event, the recursive algorithm uses transition probabilities governed by the retention rate at the event, as described by Rabier et al. (2014). $\mathbb{P}\{\text{family } i \text{ is retained}\}$ is dependent on the filtering decision. For instance, if all gene families retained for analysis are those with at least one gene copy, then $\mathbb{P}\{\text{family } i \text{ is retained}\} = \mathbb{P}\{D_i \neq (0, \ldots, 0)\}$, which is calculated recursively on the species tree from the tips to the root as before (Rabier et al. 2014). Here, we filtered gene count data to retain families such that at least one gene copy is present in the subtrees left and right of the root of the species tree. This filtering is reflected in the calculation of $\mathbb{P}\{\text{family } i \text{ isretained}\}$, again using a recursive algorithm. Thus, the likelihood function and MLEs of λ, μ, and q depend only the timing of a WGD on a fixed edge e, as well as filtering gene count data to span the root of the species tree.

After convergence of the likelihood scores for all runs, we performed a series of LRTs to determine the significance of

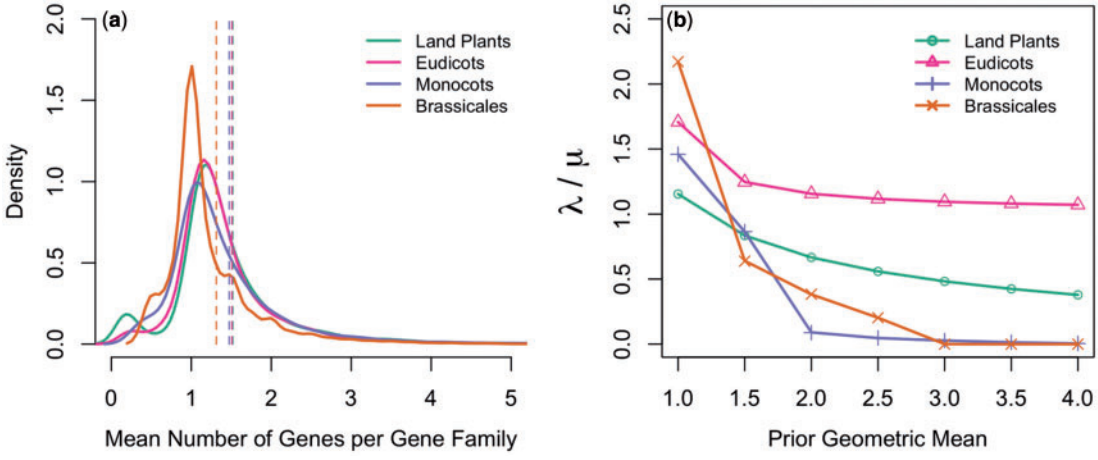

Fig. 2.—(*a*) The distribution of mean number of genes per gene family across the 22 taxa used in this study is shown using kernel densities for the land plant, eudicot, monocot, and Brassicales filtered data sets. The mean of each distribution is shown by a vertical dashed line. These are all close to 1.5, which provides some justification of a geometric mean of 1.5 as the best prior choice. (*b*) The prior used for each analysis is plotted against the ratio of the estimated duplication and loss rates. A prior of one enforces a single gene at the root of the species tree for likelihood calculations, but this can lead to inflated duplication rates. Allowing a prior geometric mean of 1.5 allows for some uncertainty in the number of genes at the root, without driving duplication rates to 0 or inflating loss rates, which happens when using higher priors.

individual putative WGD events within the land plant phylogeny. Models were nested by removing only one WGD or WGT at a time and comparing this to the model with all WGD or WGT events. This was done to avoid WGD events of large effect from influencing λ and μ, creating dependence on the order WGD events were tested. Probabilities of likelihood ratio test (LRT) statistics based on mixture densities are given in Rabier et al. (2014). A nominal probability of a type I error of 0.001, such that a significant LRT statistic must be >9.55, was applied to all tests. WGD retention rates were based on the final model, where all putative ancient WGDs are in the tree.

Exploring the Timing of Ancient WGD Events

While there is uncertainty in the precise timing of all ancient WGDs, we explored more thoroughly the timing of the putative WGDs preceding the diversifications of angiosperms and seed plants since they reside on a very long branch (fig. 1). Specifically, we tested the angiosperm WGD and seed plant WGD on 5 Myr intervals between 148–349 Ma (angiosperm) and 350–420 Ma (seed plant). All retention rates were optimized for each run, and starting gene duplication and loss rates were provided from the land plant analyses, 0.0016 and 0.0019, respectively, to speed computation.

Testing Arabidopsis Alpha and Beta Duplication Events

In our initial analysis of the land plant data set, we failed to detect the *Arabidopsis* α WGD, but we detected the *Arabidopsis* α WGD using the Brassicales data set. This could be due to 1) power alone (because the Brassicales data set contains more gene families), 2) different rates of gene retention in Brassicales specific gene families than in gene families that span the land plant root, or 3) differences in λ and μ within the Brassicales than the rest of the land plant tree. To address power and retention rate differences among gene families, we randomly sampled without replacement 7,567 (i.e., the number of gene families in the land plant data set) of the 12,957 gene families in the Brassicales data set 500 times, and for each randomly sampled data set, we estimated λ, μ, and each *q* using a geometric mean of 1.5 as the prior distribution of the number of genes at the root. To observe if the Brassicales-specific gene families were driving the result, the Brassicales model likelihood was optimized using only the 6,843 gene families shared between the land plant data and the Brassicales data, as well as only 2,170 Brassicales data set-specific gene families (i.e., families with no sequence outside of Brassicales and *T. cacao* in our tree) and only the 10,489 angiosperm specific gene families (i.e., families with no sequences outside of angiosperms). To speed up optimization for these experiments and make them computationally tractable, we used the estimates of λ and μ from the original Brassicales data set as starting values.

To address the effects of λ and μ on the *q* values of *Arabidopsis* α and β and on the LRT statistics for these two WGD events, we calculated the likelihood score of each randomly resampled Brassicales data set in several ways. First, for each of the 500 resampled data sets, λ, μ, and retention rates for both the *Arabidopsis* α and β WGDs were unconstrained and optimized by maximum likelihood. Next, we calculated the likelihood score by fixing *q* for both the *Arabidopsis* α and β WGDs to their previously optimized values, and fixing λ and μ to the estimates from the analysis of the land plant data set, 0.0016 and 0.0019, respectively. LRTs were calculated by removing either the *Arabidopsis* α or the *Arabidopsis* β WGD event. If a WGD is not statistically significant using the land plant λ and μ, then it indicates that gene duplication and loss rates estimated from the land plant tree do not adequately describe gene duplication and loss rates in the Brassicales. Using the land plant estimates of λ and μ allowed us to observe the effects of enforcing inappropriate gene duplication and loss rates on our ability to detect WGDs.

Testing Multiple Duplications on a Single Branch

Testing the grass ρ and σ as well as the *M. acuminata* α, β, and γ WGD events required further attention. For instance, since the estimation of *q* depends on the timing of the WGD event, a model without grass ρ is not equivalent to σ. Therefore, we had to explore and test all possible ways that WGD events can be removed from a branch. For testing *M. acuminata* α, β, and γ, all nine nested hypotheses were explored (fig. 3). For example, the *M. acuminata* branch with both the α and β WGDs is preferable to just α; however, having α, β, and γ improves the likelihood score further, which is equivalent to *M. acuminata* with γ alone.

Simulation Experiments

We extended the WGDgc R package (version 1.2) to simulate gene count data along a phylogeny. Simulations were used to estimate our power to detect a WGT or two WGDs on a single branch. Specifically, we ran experiments to test four sets of WGDs or WGTs: 1) the eudicot WGT, 2) *Arabidopsis* α and β, 3) the monocot WGD, *Ph. dactylifera* WGD, Poaceae ρ WGD, and Poaceae σ WGD, and 4) two WGDs predating the divergence of *A. trichopoda* and other angiosperms (i.e., the angiosperm and seed plant WGDs). All simulation experiments were performed with four-taxon trees to make them computationally feasible. The four-taxon trees used in the simulation experiments are shown in figure 4, and newick and simmap formatted trees with branch lengths are provided in the supplementary material, Supplementary Material online.

The eudicot WGT simulations were performed for the duplication and loss rates estimated from the eudicot data (λ = 0.0022, μ = 0.0018). Likewise, the *Arabidopsis* α and β WGDs were simulated using the gene duplication and loss rates estimated from the Brassicales data (λ = 0.00135, μ = 0.00211), monocot lineage WGDs were simulated under the monocot gene duplication and loss rates (λ = 0.00249,

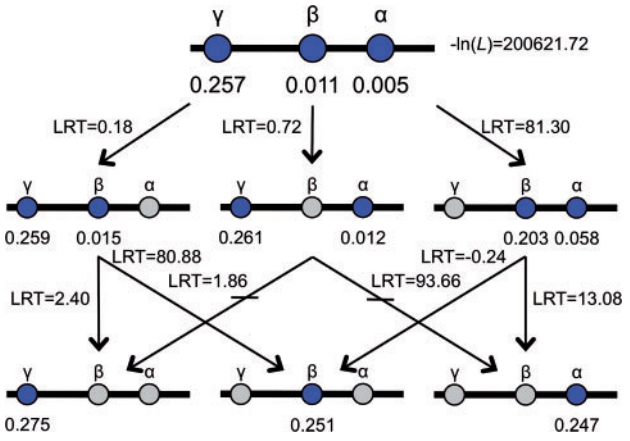

Fig. 3.—Testing multiple WGD hypotheses on a single branch using results for *Musa acuminata* α, β, and γ for the land plant data set. Blue circles represent WGDs present in the model and gray circles indicate absence. Although *M. acuminata* α and β is preferable to a model with only *M. acuminata* β, the model with all three WGDs (top) is statistically equivalent to a model with only *M. acuminata* γ (bottom left). Retention rates for each WGD for each model are displayed below each event.

$\mu = 0.00288$), and the angiosperm and seed plant WGDs were only simulated using the land plant duplication and loss rates ($\lambda = 0.00162$, $\mu = 0.00196$). The eudicot WGT simulations tested q of 0.01 or 0.10 such that either all WGDs/WGTs had equal retention rates or the eudicot WGT had $q = 0.01$, while more recent WGDs/WGTs had $q = 0.10$. We allowed unequal retention rates in the *Arabidopsis* α and β simulations, such that $q_\alpha = q_\beta = 0.01$, $q_\alpha = q_\beta = 0.10$, or $q_\alpha = 0.10$ and $q_\beta = 0.01$. Because the monocot simulations were more complex than other simulation scenarios, we only allowed all WGDs to have $q = 0.01$ or $q = 0.10$. The angiosperm and seed plant WGDs were simulated only with equal retention rates of $q_{\text{angiosperm}} = q_{\text{seed plant}} = 0.01$ or $q_{\text{angiosperm}} = q_{\text{seed plant}} = 0.10$.

For each experiment, we ran 100 simulation replicates with 500, 1,000, 5,000, and 10,000 simulated gene families. All simulated data used a geometric distribution of genes at the root of each gene family, with a mean of 1.5. Data were conditioned to span the root of the species trees used to test the eudicot WGT, the *Arabidopsis* α and β WGDs, and the angiosperm and seed plant WGDs, respectively. The trees and R scripts used to run simulations are available as supplementary material, Supplementary Material online.

Results

Evidence for Ancient WGDs in Plants

Using the full land plant data, we detected nine WGDs from of the 19 ancient WGD hypotheses tested (fig. 1). We detected WGD and WGT events with high q values relative to the estimated λ and μ consistently across all of our analyses (fig. 1; supplementary tables S1–S4, Supplementary Material online). For example, we detected WGDs that occurred in the *P. patens* lineage ($q = 0.098$) and the *Aq. coerulea* lineage ($q = 0.073$), both of which are also supported by d*S* distributions (Rensing et al. 2007; Cui et al. 2006). We also detected evidence for a WGT that is shared by *S. lycopersicum* and *S. tuberosum* ($q = 0.102$), a WGD preceding *Po. trichocarpa* ($q = 0.597$), a WGD shared by Papilionoids ($q = 0.163$), and a WGD that was specific to *G. max* ($q = 0.763$), all of which have some evidence based on synteny (Tuskan et al. 2006; The Tomato Genome Consortium 2012; Schmutz et al. 2010). We detected the oldest of three proposed WGDs that occurred in the *M. acuminata* lineage ($q = 0.258$) as well as the older of two proposed events that preceded the diversification of Poaceae ($q = 0.062$). We also detected the older WGD in the *Ar. thaliana* lineage, *Arabidopsis* β WGD ($q = 0.063$), using the full land plant data set, but we detected both *Arabidopsis* α ($q = 0.023$) and β ($q = 0.088$) using *T. cacao* and the Brassicales only (i.e., the Brassicales data set; fig. 1).

The estimates of the proportion of genes retained from WGDs varied greatly among putative ancient WGD events (fig. 1). These retention rates were robust to both the choice of prior and the specific gene set being used (supplementary tables S1–S4, Supplementary Material online). The prior in these analyses is the probability distribution of the number of gene copies at the root of the species tree, which accounts for uncertainty in ancestral gene family sizes. A geometric mean of 1.5 is a reasonable prior choice based on the distribution of average gene family sizes (fig. 2*a*). Prior choice had little effect on estimates of λ and μ in the land plant analyses, but it affected our ability to detect some WGD or WGT events in data sets that span the root of *A. trichopoda* and the eudicots (the eudicot data set) as well as *T. cacao* and the Brassicales (the Brassicales data set; fig. 2*b*; supplementary tables S2 and S4, Supplementary Material online). Specifically, we did not detect the WGT preceding *Solanum* when we assumed the number of genes at the root of *A. trichopoda* and the eudicots was geometrically distributed with a mean of one (i.e., only one gene copy at the root of the tree for each gene family; supplementary table S2, Supplementary Material online). An inappropriately high prior for the analyses of *T. cacao* and Brassicales led to a failure to detect both the *Arabidopsis* α and β WGDs, but we detected both events when we assume the number of genes at the root is geometrically distributed with a mean of 1.5 (supplementary table S4, Supplementary Material online).

Challenges in Testing WGD Hypotheses

In several cases, our analyses did not detect multiple WGDs on the same branch on a phylogeny. Tests of the three consecutive WGDs on the branch leading to *M. acuminata* (*M. acuminata* α, β, and γ) favored a scenario with only the oldest (γ) WGD. Placing a single WGD close to 96 Ma on the *M.*

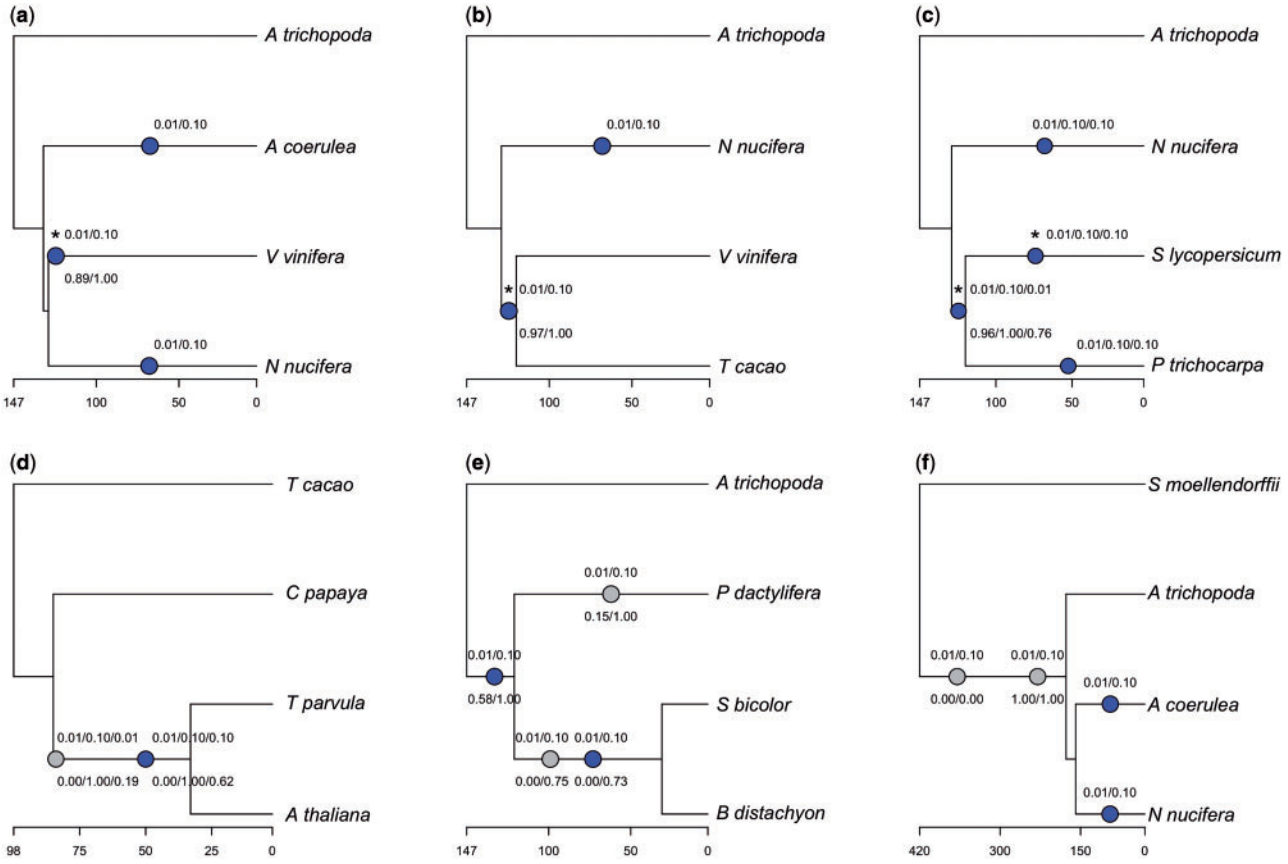

Fig. 4.—Gene count data were simulated on four-taxon trees to test WGD hypotheses and estimate power. Events associated with significant LRTs in the observed data are in blue while nonsignificant events are colored gray. An "*" above a circle indicates a WGT rather than a WGD. Numbers above circles indicate the simulated retention rates and numbers below circles indicate power. Panels (*a*)–(*c*) were designed to test the eudicot WGT. Panels (*d*)–(*f*) were designed to test two WGDs on a single branch. Panel (*d*) estimates power for testing the *Arabidopsis* α and β WGDs on the same branch. Panel (*e*) simulates a complex scenario of WGDs in the monocot lineage for the monocot WGD, *Phoenix dactylifera* WGD, Poaceae ρ WGD, and Poacea σ WGD. Panel (*f*) is used to assess our ability to detect the angiosperm and seed plant WGDs.

acuminata branch resulted in a significant improvement in the likelihood score (fig. 3). LRTs showed no statistical support for WGDs in *M. acuminata* at 65 Ma (β) and 60 Ma (α). When testing these WGDs with data that include only the monocots and *A. trichopoda*, a single WGD on the branch of *M. acuminata* near 96 Ma, with $q = 0.275$ was also preferred.

Testing the two WGD hypotheses before the diversification of Poaceae with the land plant data set also suggested only a single WGD near 96 Ma (Poaceae σ; fig. 1). However, we did not detect the Poaceae σ WGD with the monocot data set (supplementary table S3, Supplementary Material online). The more recent (70 Ma) Poaceae ρ WGD was not statistically significant and had estimated retention rates of 0.001 and 0 using both the land plant (fig. 1; supplementary table S1, Supplementary Material online) and monocot data sets (supplementary table S3, Supplementary Material online), respectively.

We detected the *Arabidopsis* β WGD based on the hypothesis in figure 1, where we placed the *Arabidopsis* α and β WGDs are placed on separate branches. However, we did not detect the *Arabidopsis* β WGD when both *Arabidopsis* α and β WGDs were on the same branch (supplementary table S5, Supplementary Material online). In addition, we did not detect evidence of the most ancient WGDs, including those preceding the diversification of angiosperms and seed plants (fig. 1). We explored the timing of the angiosperm and seed plant WGDs by adjusting the age of the putative WGDs along 5 Ma intervals on the branch prior to the divergence between *A. trichopoda* and other angiosperms, and in all cases we were still unable to detect the angiosperm and seed plant WGDs (supplementary table S6, Supplementary Material online). We also did not find evidence of a WGT common to all eudicots in both the land plant and the eudicot data sets (fig. 1; supplementary tables S1 and S2, Supplementary Material online); however, we detected the eudicot WGT using four-taxon trees (see Evaluating Model Performance with Simulations). A WGD common to all monocots had a small but insignificant $q = 0.007$ in the analysis of the land

plant data set and $q = 0$ in the monocot data set (fig. 1; supplementary tables S1 and S3, Supplementary Material online).

Evaluating Model Performance with Simulations

We performed simulation experiments to test model performance in cases where we did not detect a well-documented ancient WGD and determine if the failure to detect a WGD or WGT was due to a lack of power. For the eudicot WGT, we tested three simulation scenarios: 1) a WGT on a terminal branch represented by *V. vinifera* (fig. 4*a*), 2) a WGT on an internal branch, without any WGDs following the eudicot WGT (fig. 4*b*), and 3) a WGT on an internal branch with a WGD on the tip leading to *Po. trichocarpa* and WGT on the tip leading to *S. lycopersicum* after the eudicot WGT (fig. 4*c*). For testing the eudicot WGT on a single terminal branch (fig. 4*a*), we had power of 0.76 to detect the WGT when $q = 0.01$ with 5,000 gene families (supplementary table S7, Supplementary Material online) and power of 0.89 with 10,000 gene families. Similarly, when the eudicot WGT is on an internal branch with no other WGDs or WGTs on the tips (fig. 4*b*), power = 1 when $q = 0.10$ and there were at least 500 gene families, and power = 0.97 for $q = 0.01$ when there were 10,000 gene families (supplementary table S8, Supplementary Material online). When the eudicot WGT was placed on an internal branch with an additional WGT and WGD occurring on terminal branches (fig. 4*c*), we had power of 0.96 and 1 for 10,000 gene families when $q = 0.01$ and $q = 0.10$, respectively, for all WGD and WGT events (fig. 4*c*; supplementary table S9, Supplementary Material online). However, when the eudicot WGT had $q = 0.01$ and the more recent events on terminal branches had $q = 0.10$, we have power of 0.76 to detect the eudicot WGT for 10,000 gene families (fig. 4*c*; supplementary table S9, Supplementary Material online). In contrast to the analyses using the plant data and the full tree, we detected the eudicot WGT in the observed data using the three four-taxon trees. The first case, the WGT occurs on the *V. vinifera* branch without any consecutive WGDs or WGTs (fig. 4*a*; supplementary table S7, Supplementary Material online). Second, the WGT occurs before the divergence of *V. vinifera* and *T. cacao* and is not obstructed by any other WGT or WGD on the tree (fig. 4*b*; supplementary table S8, Supplementary Material online). In the third case, the eudicot WGT occurred before the divergence of *Po. trichocarpa* and *S. lycopersicum*; afterward, there is also a WGD in the tip leading to *Po. trichocarpa* and a WGT in the tip leading to *S. lycopersicum* (fig. 4*c*; supplementary table S9, Supplementary Material online). The eudicot WGT had an estimated q of 0.070, 0.067, and 0.104 for these analyses, respectively.

Simulation results suggest that detecting multiple WGDs on a single branch is difficult, if not impossible, when retention rates are low (supplementary table S10, Supplementary Material online). Specifically, we had no power to detect both the *Arabidopsis* α and β WGDs when $q_\alpha = q_\beta = 0.01$, even with 10,000 gene families (fig. 4*d*). However, the power to detect both the *Arabidopsis* α and β WGDs goes to 1 when $q_\alpha = q_\beta = 0.10$ for 10,000 gene families. When we simulated gene count data with $q_\alpha = 0.10$ and $q_\beta = 0.01$, we had power of 0.02–0.62 and 0.04–0.19 to detect the *Arabidopsis* α and β WGDs, respectively, for 500–10,000 gene families. When testing the observed data on the same phylogeny used for simulating the *Arabidopsis* α and β WGDs on the same branch, we did not detect the *Arabidopsis* β WGD, but we detected the *Arabidopsis* α WGD with an estimated $q = 0.092$ (supplementary tables S5 and S10, Supplementary Material online). This is consistent with our simulation results, which suggests that with 10,000 gene families, we had power of 0.62 to detect the *Arabidopsis* α WGD when $q_\alpha = 0.10$ and only power of 0.19 to detect the *Arabidopsis* β WGD for $q_\beta = 0.01$ (fig. 4*d*; supplementary table S10, Supplementary Material online).

Testing the WGD hypotheses associated with monocots presents a complex scenario with nested WGD hypotheses (fig. 4*e*). For simulated data with $q = 0.01$ for all WGD events, we had no power to detect both the Poaceae ρ and the Poaceae σ WGDs, power of 0.04–0.15 to detect the *Ph. dactylifera* WGD, and power of 0.05–0.58 to detect the monocot WGD for 500–10,000 gene families (supplementary table S11, Supplementary Material online). However, when $q = 0.10$ for all the monocot WGDs, we had power of 0.73, 0.75, 1, and 1 to detect the Poaceae ρ, Poaceae σ, *Ph. dactylifera*, and monocot WGDs, respectively, with 10,000 gene families (supplementary table S11, Supplementary Material online). When testing the 10,795 monocot gene families on the four-taxon tree used for simulation (fig. 4*e*), we detected the monocot WGD with estimated $q = 0.309$ and the Poaceae ρ WGD with estimated $q = 0.053$; however, we did not detect the *Ph. dactylifera* WGD or the Poaceae σ WGD (supplementary table S11, Supplementary Material online). Notably analyses of the larger phylogenetic hypotheses with land plant data (fig. 1; supplementary table S1, Supplementary Material online) and monocot data (supplementary table S3, Supplementary Material online) did not detect the putative ancient WGD preceding the diversification of monocots but analysis of the four-taxon tree (fig. 4*e*) detected the WGD (supplementary table S11, Supplementary Material online).

We also simulated data to address the WGDs preceding the diversification of seed plants and angiosperms. Because of the available whole-genome sampling, the angiosperm and seed plant WGDs were located on the same branch, and they were the oldest WGDs tested in this study (fig. 4*f*). Simulations indicated that we had power of 0.09, 0.41, and 1 to detect the angiosperm WGD for 500, 1,000, and 5,000 or more gene families when $q = 0.01$ (supplementary table S12, Supplementary Material online). However, we had power of 1 to detect the angiosperm WGD for $q = 0.10$ when there were at least 500 gene families (supplementary table S12, Supplementary Material online). In the same simulations, we

had no power (i.e., power = 0) to detect the seed plant WGD regardless of *q* or the number of gene families (supplementary table S12, Supplementary Material online). We did not detect either the angiosperm or seed plant WGD when testing the observed land plant data on the four-taxon tree used for simulations (fig. 4*f*; supplementary table S12, Supplementary Material online).

In the four simulation experiments that included *N. nucifera* in the four-taxon tree, we detected the WGD preceding *N. nucifera* with *q* ranging from 0.171 to 0.250 (fig. 4; supplementary tables S7–S9 and S12, Supplementary Material online).

Heterogeneity in Gene Duplication and Loss across Lineages and Gene Families

The *Arabidopsis* α WGD was not significant in analyses using the land plant or eudicot data sets; however, it was detected when testing WGDs with the four-taxon Brassicales data set (fig. 1). Using the four-taxon tree requires a different gene count data set to span the root; this includes 12,967 gene families, 6,843 of which are shared with the land plant data set. Depending on the root of the tree and the conditioning of the data, different gene families can be considered in the analysis (fig. 1). Apart from simulations, we tested if failure to detect the *Arabidopsis* α WGD when using the land plant data set was due to 1) lower power, because there were fewer genes in the land plant data set than in the Brassicales data set, 2) a lack of signal in older gene families as opposed to de novo gene families in the Brassicales data set or angiosperm data, or 3) poor estimates of the single gene duplication and loss rates (λ and μ) in Brassicales when using λ and μ estimates from across the entire land plant tree.

We randomly resampled without replacement 7,567 gene families from the Brassicales data set (i.e., the number of gene families in the land plant data set) 500 times to test if the land plant data lacked power (see Materials and Methods). When we estimated λ and μ for the 500 randomly resampled Brassicales data sets (supplementary table S13, Supplementary Material online), the *Arabidopsis* α WGD was always significant (fig. 5*a*). Therefore, a lack of power does not explain why we fail to detect the *Arabidopsis* α WGD with the land plant data set. Additionally, the *Arabidopsis* α was detected when using only the 6,843 gene families that span the root of both the land plant tree and the four-taxon tree. Thus, Brassicales de novo genes alone cannot explain the significant *Arabidopsis* α LRT in the four-taxon tree.

Since the *Arabidopsis* α WGD was detected on the four-taxon tree using the 6,843 gene families in common between the land plant data and Brassicales data, the estimates of λ and μ from the land plant tree may not be appropriate for optimizing the retention rate of the *Arabidopsis* α WGD. Specifically, the global gene duplication rate was lower and the global gene loss rate was higher in the Brassicales data set (λ = 0.0013 and μ = 0.0021) than the global estimates from the full land plant tree (λ = 0.0016 and μ = 0.0019). Calculating the likelihoods based on the retention rates from the four-taxon Brassicales tree but using estimates of λ and μ based on the land plant tree (See Materials and Methods for details) led to a nonsignificant LRT for the analysis using the 6,843 gene families that span both land plants as well as *T. cacao* and Brassicales in addition to nine out of the 500 resampled data sets (fig. 5*b*). Therefore, inappropriate estimates of λ and μ and a stronger signal for WGD from more recent gene families contributed to the failure to detect the *Arabidopsis* α WGD in the land plant data set. Estimates for *q* from the *Arabidopsis* α WGD on the four-taxon tree using land plant data, angiosperm data, and Brassicales-only data were 0.023, 0.027, and 0.043, respectively (fig. 5*a*).

The *Arabidopsis* β WGD was significant in all analyses of the resampled data sets as well as analyses using gene families shared by land plants and Brassicales as well as angiosperms and Brassicales but not Brassicales alone (fig. 5*c*). When optimizing λ and μ with the Brassicales specific gene families, the *Arabidopsis* β WGD had an estimated *q* of 0. The *Arabidopsis* β WGD was significant when fixing λ and μ to rates obtained from the land plant analysis for all data sets (the 500 resampled Brassicales data sets, the land plant data set, the angiosperm data set, and the Brassicales-only data set) using the previous estimates of *q* (fig. 5*d*). Therefore, evidence for the *Arabidopsis* β WGD was stronger in older gene families that were better characterized by λ and μ optimized from the global land plant data than more recent gene families that had a higher estimated λ. Retention rates for the *Arabidopsis* β WGD on the four-taxon tree using land plant data, angiosperm data, and Brassicales-only data were 0.123, 0.026, and 0.027, respectively.

Discussion

The growth of genomic data in plants has prompted much interest in identifying evidence of ancient WGDs and associating ancient WGD with diversification and trait evolution (e.g., Vanneste et al. 2014; Cannon et al. 2015). Our analyses suggest that in many cases we can evaluate ancient WGD hypotheses and assess the impact of ancient WGDs on gene content using only gene count data. The gene count model of Rabier et al. (2014) provides an effective and statistically rigorous test for ancient WGDs in plants, while also providing estimates of gene retention rates following individual WGD events. This model does not require sequences to be assembled at the chromosome level, or dating based on d*S*, but only annotation of genome sequence data and circumscribed gene families.

The WGD retention rates (*q*) following WGDs vary tremendously across the tree (fig. 1). While a few WGDs, including those associated with *Po. trichocarpa* and *G. max* appear to

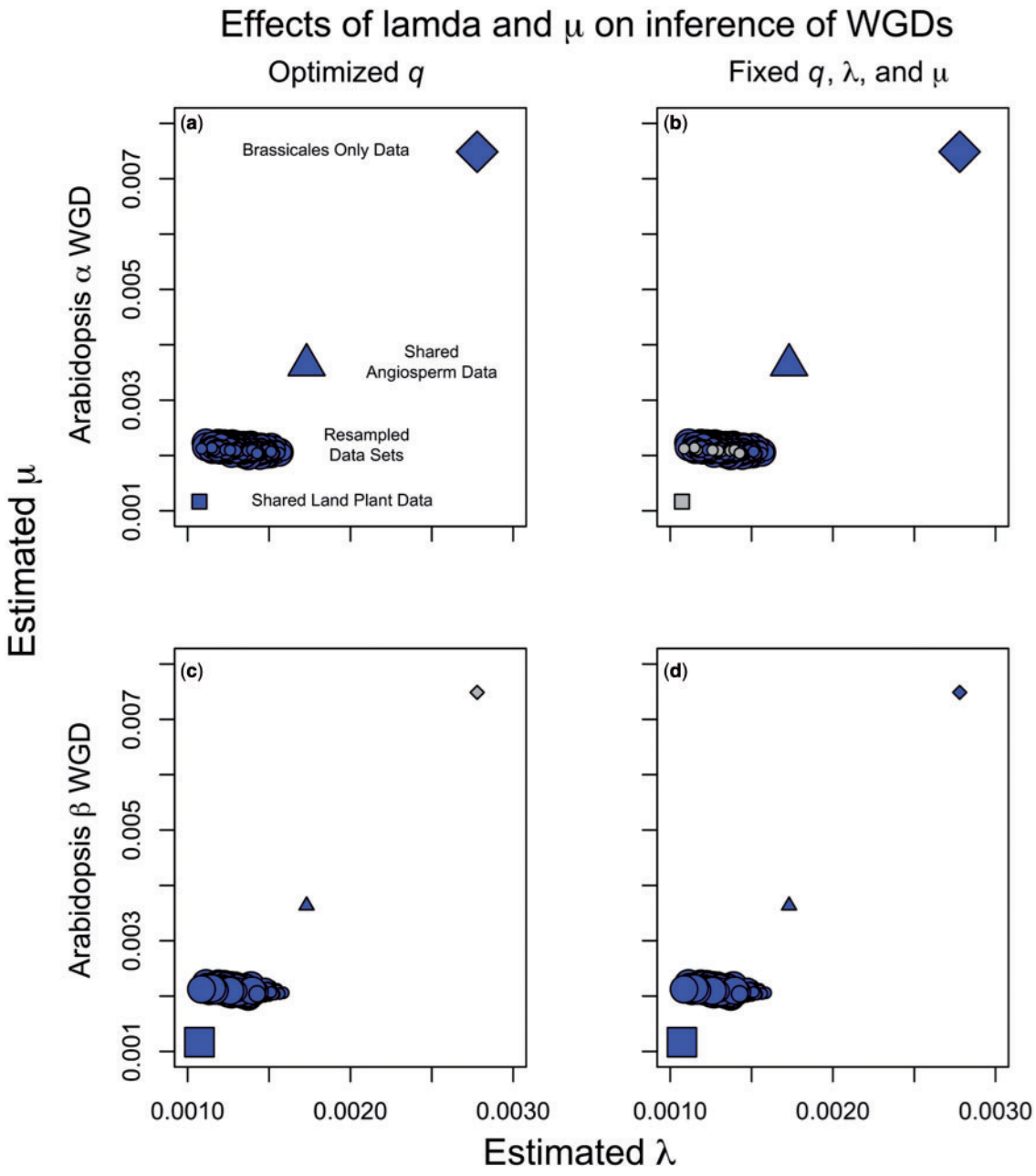

Fig. 5.—Circles represent data sets of 7,564 genes randomly sampled without replacement from the Brassicales data set, while the square represents the land plant data set with 7,564 gene families that also span the root of Brassicales and *Theobroma cacao*. The triangle represents gene families that span the root of angiosperms as well as Brassicales and *T. cacao*, and the diamond represents gene families that only span the root of Brassicales and *T. cacao*. The size of each point represents retention rate quartiles. A significant LRT statistic is colored blue and nonsignificant LRTs are gray. Panels (*a*) and (*c*) show the estimated λ, μ, and *q* for the *Arabidopsis* α and β WGDs, respectively, as well as the result of the LRT statistic. In panels (*b*) and (*d*), the likelihoods were calculated by fixing the retention rates to their values optimized in panels (*a*) and (*c*) and by further fixing λ and μ to 0.0016 and 0.0019 from the land plant data. Data points are plotted according to λ and μ estimated in panels (*a*) and (*b*) to separate the points and to better show effects of inappropriate parameter estimates on testing WGD hypotheses with gene count data.

retain a majority of the duplicated genes, only a small percentage of duplicated genes survive most ancient WGDs (fig. 1). Some of the variation in retention rates may be due to the type of WGD event, and consequently, whether fractionation is biased or unbiased (Freeling 2009). During fractionation, there may be biases regarding the types of genes that are retained following WGD (Thomas et al. 2006; Li et al. 2016) as well as which genome copy's genes are retained (Schnable

et al. 2011). Unbiased fractionation suggests that gene expression and deletion affects both parental genomes equally. Recent data support unbiased fractionation in *G. max*, *Po. trichocarpa*, and *M. acuminata*, possibly indicating that these WGDs represent autopolyploidy rather than allopolyploidy (Garsmeur et al. 2014). Interestingly, in our analyses, these lineages have WGDs with the highest *q* estimates (fig. 1).

Our analyses also suggest several reasons to be cautious about identifying ancient WGDs from gene count data. First, the model of Rabier et al. (2014) assumes that the global gene duplication and loss rates (λ and μ) are constant throughout the tree. When testing WGDs on a tree spanning a large phylogenetic distance, λ and μ will be averaged across lineages, and this may lead to inappropriate estimates of λ and μ. For example, our analyses suggest that λ and μ are different in the four-taxon tree (Brassicales and *T. cacao*) than in other parts of the land plant tree, and this affects tests of the *Arabidopsis* α WGD (fig. 1). Although most WGDs we tested, such as *Arabidopsis* β, appear to be relatively robust to poor estimates of these parameters, tests of WGDs on more limited phylogenies, or alternatively, tests allowing different λ and μ values in different parts of the tree may improve the accuracy of the analyses. However, enabling local λ and μ values on branches throughout the tree also may diminish the power to identify WGDs. For example, it may be difficult to distinguish a WGD on a branch from an increased λ on the same branch. The gene count model also implicitly assumes that all gene families have the same λ and μ and that all genes have equal probability of being retained following WGDs. There is strong evidence that gene retention after fractionation often is not random (Blanc and Wolfe 2004b; Seoighe and Gehring 2004; Maere et al. 2005; Rizzon et al. 2006; Makino and McLysaght 2012; Li et al. 2016) or that fractionation is not necessarily complete before a following speciation event (Schranz et al. 2012; Conant 2014).

In our study, the retention rates of the *Arabidopsis* α and β WGDs for gene families that only span the root of Brassicales and *T. cacao* were higher for the *Arabidopsis* α WGD and lower for the *Arabidopsis* β WGD than the retention rates for older gene families that span the root of the land plant tree. These inconsistent retention rate estimates suggest the possibility that a higher proportion of gene families that span the root of land plants were retained by the *Arabidopsis* β WGD as compared to the more recent *Arabidopsis* α WGD. This may explain why we detected this event with the land plant data set but not the four-taxon analysis where both *Arabidopsis* α and β WGDs are on the same branch. Another explanation is that older WGDs and WGTs, such as the eudicot WGT, had not completed the fractionation process prior to the *Arabidopsis* β WGD.

A critical remaining question is whether our failure to detect several putative ancient WGDs in any analyses (fig. 1) is due to the model or lack of power for the gene count method or whether it suggests that the putative ancient WGDs did not happen. In 9 out of the 19 putative ancient WGDs or WGTs tested in this study, including the WGT prior to the diversification of eudicots, we did not detect a WGD or WGT in analyses using the whole land plant tree (fig. 1). However, we detected the eudicot WGT, the *Arabidopsis* α WGD, and the *Nelumbo* WGD using four-taxon trees. There are multiple lines of evidence for the eudicot WGT, including syntenic comparisons between *V. vinifera* and other angiosperms (Jaillon et al. 2007; Argout et al. 2011) and comparisons between *V. vinifera* and *A. trichopoda* (*Amborella* Genome Project 2013). For example, a 1:3 ratio of syntenic blocks between *A. trichopoda* and *V. vinifera* would suggest a WGT occurred after the divergence of *A. trichopoda* and other angiosperms but before the divergence of eudicots (*Amborella* Genome Project 2013). Our ability to detect the eudicot WGT in analyses using the whole tree may be obscured by the presence of more recent WGDs and heterogeneity in retention rates across WGDs and WGTs when analyzing large phylogenies. Therefore, there may be benefits to refining hypothesis tests with smaller phylogenies, since we do detect the eudicot WGT in all three four-taxon tree hypotheses (fig. 4*a*–*c*). Our simulation experiments also suggest it is possible that we did not detect the eudicot WGT in the full land plant (fig.1; supplementary table S1, Supplementary Material online) or eudicot (supplementary table S2, Supplementary Material online) analyses because the eudicot WGT has a low retention rate in comparison to more recent WGD and WGT events (fig. 4*c*, supplementary table S9, Supplementary Material online). We might expect older WGDs to have more time for gene loss; therefore, it may be generally more difficult to detect WGDs and WGTs with increasing age. In any case, our results indicate that the retention rate associated with the eudicot WGT is low, which is consistent with previous analyses of gene family evolution (*Amborella* Genome Project 2013), and consequently, it is difficult to detect, especially with the complexities of testing nested WGDs and WGTs.

Simulations indicate that the gene count model lacks power to detect multiple WGDs on a single branch, especially when more recent WGDs have a higher retention rate compared to older WGDs (supplementary tables S5 and S10–S12, Supplementary Material online). Therefore, some of our results should be interpreted with caution: an absence of evidence for a WGD is not evidence of the absence for a WGD. We resolved the difficulty of identifying both the *Arabidopsis* α and β WGDs by placing the *Arabidopsis* α WGD on a separate, neighboring branch. Although recent evidence suggests that this placement is not correct (Dassanyake et al. 2011; Vanneste et al. 2014), this enables us to detect both the *Arabidopsis* α and β WGDs. Ideally it would be possible to include lineages that break up successive WGD events onto separate branches when testing WGD hypotheses.

Evidence for some putative ancient WGDs that we did not detect is largely circumstantial. For example, the WGD on the

terminal branch leading to *Ph. dactylifera* was proposed based on the distribution of d*S* between paralogs and some syntenic data for the largest scaffolds of the *Ph. dactylifera* genome sequence (Al-Mssallem et al. 2013). Our simulations suggest we have should have detected this WGD with 10,000 gene families if $q = 0.1$, but we only have power of 0.15 to detect the *Ph. dactylifera* WGD if $q = 0.01$. Thus, our results suggest that either the retention rate for the *Ph. dactylifera* WGD is extremely low or that there is no ancient WGD in the lineage leading to *Ph. dactylifera*.

The hypothesis that two WGDs occurred before the diversification of grasses (i.e., Poaceae ρ and σ) is based on 2,416 genes on nine syntenic blocks in the *O. sativa* genome and 831 genes on eight older syntenic blocks, inferred by ratios of duplicate genes, respectively (Simillion et al. 2002; Tang et al. 2010). Although genes on the Poaceae ρ and σ correspond to different median d*S* values (Tang et al. 2010), our results imply that only one WGD occurred before the divergence of grasses. The different *dS* values could be due to paralogs in genomic regions where synteny has eroded over time having higher substitution rates than paralogs maintained on syntenic blocks that are detectable by comparison of *O. sativa* chromosomes and not two separate WGDs. However, in the case of Poaceae ρ and σ, our simulations indicate we have at most power or 0.73 and 0.75 to detect the two WGDs assuming $q = 0.1$, and if $q = 0.01$, we have no power to detect both WGDs. Thus, although evidence for two WGDs preceding the diversification of Poaceae is far from conclusive, we cannot rule out the possibility that they occurred but have extremely low retention rates.

Similarly, there is some syntenic evidence for a WGD prior to the diversification of monocots (e.g., Tang et al. 2010); however, complexities of testing the monocot WGD prevent us from differentiating model performance and biology. We did not detect the monocot WGD in the observed data on the full land plant data set (fig. 1; supplementary table S3, Supplementary Material online). However, simulations suggest we have excellent power to detect the monocot WGD on the four-taxon tree when $q = 0.1$ and power of 0.58 for $q = 0.01$ with 10,000 gene families (fig. 4*e*; supplementary table S11, Supplementary Material online). We detected the monocot WGD on the four-taxon tree with the observed data (fig. 4*e*; supplementary table S11, Supplementary Material online); however, the estimated q of the monocot WGD was 0.309. Based on other analyses of ancient WGDs with plant genomes (e.g., Tang et al. 2010), a retention rate of 0.309 is higher than expected, and it may be an artifact from testing many nested WGD hypotheses (fig. 5*e*). Extensive testing revealed this result is not due to optimization error. Instead it may be due to an excess of duplicate genes in *Sor. bicolor* compared to the other taxa in the four-taxon tree.

We also find no evidence for WGDs that predated the divergence of angiosperms or seed plants based on the gene content of current genomes. This is not surprising, as there is only weak syntenic support for the angiosperm WGD and none for the seed plant WGD (*Amborella* Genome Project 2013). Furthermore, the distribution of d*S* based on a few hundred paralogs may be unreliable at such an old age, especially since substitutional saturation alone can lead to false positives for WGDs (Vanneste et al. 2013). Our simulations indicate that even with a $q = 0.01$, we should have sufficient power to detect the angiosperm WGD with 5,000–10,000 gene families on the four-taxon tree (fig. 4*f*; supplementary table S12, Supplementary Material online). In contrast, we have no power to detect the seed plant WGD, even if q is as high as 0.1 (fig. 4*f*; supplementary table S12, Supplementary Material online). Thus, although again we cannot rule out an ancient WGD preceding the angiosperms if $q < 0.01$, our analyses provide some reasons to question the angiosperm WGD. Because of lack of power, we cannot evaluate the putative seed plant WGD based on gene count data, suggesting it may be difficult to find conclusive evidence for such extremely ancient WGDs.

Additional experiments with the *Arabidopsis* α and β WGDs indicate that gene families have different probabilities of being retained following WGD (Freeling 2009, Li et al. 2016). Although we did not consider differences in gene functions or gene dosage balance (Edger and Pires 2009), we used different data filtering strategies and gene family age to show differential background duplication and loss rates (λ and μ) as well as probabilities of retention following WGD across gene families. While we primarily use the gene count model as a means of detecting WGDs, it is possible to use a model testing approach for more targeted hypothesis testing of gene family evolution following WGDs, where gene function or ontological categories are known.

Our results suggest the need to critically evaluate evidence for some ancient land plant WGD hypotheses, as well as the role these ancient WGDs have played in shaping plant genomes. Multiple lines of evidence, including this study, indicate that many duplicate gene copies in some genomes such as *Po. trichocarpa* and *G. max* are products of WGDs (Tuskan et al. 2006; Schmutz et al. 2010). However, our extremely low retention rate estimates, which are consistent with relatively stable gene numbers across land plants, indicate that the contributions of many ancient WGDs in terms of gene content often appear to be relatively minor. In several cases, we cannot distinguish the absence of a putative ancient WGD from its presence with an extremely low gene retention rate. While this model makes some simplifying assumptions, such as global background λ and μ, it is one of the few statistically rigorous approaches that has been used to evaluate evidence for ancient plant WGDs, and it can be used with taxa that are not completely mapped or sequenced. This approach also could be extended to taxa with only transcriptome data if the number of missing gene copies can be estimated (see Rabier et al. 2014). Characterizing ancient WGDs may be inherently difficult as information is lost with age, but our results

suggest that gene copy data can provide new insights into plant ancient WGDs.

Supplementary Material

Supplementary tables S1–S13 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgment

This work was supported by the National Science Foundation (DEB-1208428 to J.G.B.).

Literature Cited

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

Al-Mssallem IS, et al. 2013. Genome sequence of the date palm *Phoenix dactylifera*. Nat Commun. 4:2274.

Amborella Genome Project 2013. The *Amborella* genome and the evolution of flowering plants. Science 342:1241089.

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

Argout X, et al. 2011. The genome of *Theobroma cacao*. Nat Genet. 43:101–108.

Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. Am J Bot. 97:1296–1303.

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell 12:1093–1101.

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

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

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

The *Brassica rapa* Genome Sequencing Consortium 2011. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet. 43:1035–1039.

Burleigh JG. 2012. Identifying the phylogenetic context of whole-genome duplications in plants. In Soltis PS, Soltis DE, editors. Polyploidy and Genome Evolution. New York: Springer. pp. 77–92.

Cannon SB, et al. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol Biol Evol. 32:193–210.

Conant GC. 2014. Comparative genomics as a time machine: how relative dosage and metabolic requirements shaped the time-dependent resolution of yeast polyploidy. Mol Biol Evol. 31:3184–3193.

Conant GC, Wolfe KH. 2008. Probabilistic cross-species inference of orthologous genomic regions created by whole-genome duplication in yeast. Genetics 179:1681–1692.

Cui L, et al. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16:738–749.

Csűrös M, Miklós I. 2009. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol. 26:2087–2095.

Dassanayake M, et al. 2011. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet. 43:913–918.

D'Hont A, et al. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. Nature 488:213–219.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

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.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 17:368–376.

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

Garsmeur O, et al. 2014. Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol. 31:448–454.

Gu X, Wang Y, Gu J. 2002. Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet. 31:205–209.

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

Hedges SB, Dudley J, Kumar S. 2006. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics 22:2971–2972.

Jaillon O, et al. 2004. Genome duplication in the teleost fish *Tetraodon nigrovirdis* reveals the early vertebrate proto-karyotype. Nature 431:946–957.

Jaillon O, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–468.

Jiao Y, et al. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–102.

Jiao Y, et al. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 13:R3.

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

Li Z, et al. 2015. Early genome duplications in conifers and other seed plants. Sci Adv. 1:e1501084.

Li Z, et al. 2016. Gene duplicability of core genes is highly consistent across all angiosperms. Plant Cell 28:326–344.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.

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.

Lyons E, Pedersen B, Kane J, Freeling M. 2008. The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids. Tropical Plant Biol. 1:181–190.

Maere S, et al. 2005. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 102:5454–5459.

Makino T, McLysaght A. 2012. Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant. Genome Res. 22:2427–2435.

Mandáková T, Joly S, Krzywinski M, Mummenhoff K, Lysak MA. 2010. Fast diploidization in close mesopolyploid relatives of *Arabidopsis*. Plant Cell 22:2277–2290.

McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nat Genet. 31:200–204.

Ming R, et al. 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996.

Ming R, et al. 2013. Genome of the long-living scared lotus (*Nelumbo nucifera* Gaertn.). Genome Biol. 14:R41.

Otto SP. 2007. The evolutionary consequences of polyploidy. Cell 131:452–462.

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

R Development Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
Rabier C, Ta T, Ané C. 2014. Detecting and locating whole genome duplications on a phylogeny, a probabilistic approach. Mol Biol Evol. 30:750–762.
Raes J, Vandepoele K, Simillion C, Saeys Y, Van de Peer Y. 2003. Investigating ancient duplication events in the *Arabidopsis* genome. J Struct Funct Genomics 3:117–129.
Rensing SA, et al. 2007. An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. BMC Evol Biol. 7:130.
Rizzon C, Ponger L, Gaut BS. 2006. Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol. 2:e115.
Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms – inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol. 14:23.
Schmutz J, et al. 2010. Genome sequence of the paleopolyploid soybean. Nature 463:178–183.
Schnable JC, Freeling M, Lyons E. 2012. Genome-wide analysis of syntenic gene deletion in grasses. Genome Biol Evol. 4:265–277.
Schnable JC, Springer NM, Freeling M. 2011. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 108:4069–4074.
Schranz ME, Mohammadin S, Edger PP. 2012. Ancient whole genome duplications, novelty and diversification: the WGD radiation lag-time model. Curr Opin Plant Biol. 15:147–153.
Seoighe C, Gehring C. 2004. Genome duplication led to highly selective expansion of the *Arabidopsis* proteome. Trends Genet. 20:461–464.
Simillion C, Vandepoele K, Van Montagu M, Zabeau M, Van de Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 99:13627–13632.
Soltis DE, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot. 98:704–730.
Sun M, et al. 2015. Deep phylogenetic incongruence in the angiosperm clade Rosidae. Mol Phylogent Evol. 83:156–166.
Tang H, Bowers JE, Wang X, Paterson AH. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc Natl Acad Sci U S A. 107:472–477.
Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in does-sensitive genes. Genome Res. 16:934–946.
Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641.
Tuskan GA, et al. 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604.
Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci U S A. 101:1638–1643.
Vanneste K, Baele G, Maere S, Van de Peer Y. 2014. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. Genome Res. 24:1334–1347.
Vanneste K, Van de Peer Y, Maere S. 2013. Inference of genome duplications from age distributions. Mol Biol Evol. 30:177–190.
Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplication in *Arabidopsis*. Science 290:2114–2117.
Young ND, et al. 2011. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature 480:520–524.

Associate editor: Kenneth Wolfe