Widespread genome duplications throughout the history of flowering plants

Liying Cui, P. Kerr Wall, James H. Leebens-Mack, Bruce G. Lindsay, Douglas E. Soltis, Jeff J. Doyle, Pamela S. Soltis, John E. Carlson, Kathiravetpilla Arumuganathan, Abdelali Barakat, Victor A. Albert, Hong Ma and Claude W. dePamphilis

Genome Res. 2006 16: 738-749; originally published online May 15, 2006;
Access the most recent version at doi:10.1101/gr.4825606

Letter

Widespread genome duplications throughout the history of flowering plants

Liying Cui,[1,2,3] P. Kerr Wall,[1,2,3] James H. Leebens-Mack,[1,2,3] Bruce G. Lindsay,[5] Douglas E. Soltis,[6] Jeff J. Doyle,[8] Pamela S. Soltis,[7] John E. Carlson,[2,3,4] Kathiravetpilla Arumuganathan,[9] Abdelali Barakat,[1,2,3] Victor A. Albert,[10] Hong Ma,[1,2,3] and Claude W. dePamphilis[1,2,3,11]

[1]Department of Biology, [2]Institute of Molecular Evolutionary Genetics, [3]Huck Institutes of the Life Sciences, [4]School of Forest Resources, and [5]Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, USA; [6]Department of Botany and [7]Florida Museum of Natural History, University of Florida, Gainesville, Florida 32611, USA; [8]Department of Plant Biology, Cornell University, Ithaca, New York 14853, USA; [9]Virginia Mason Research Center, Benaroya Research Institute, Seattle, Washington 98101, USA; [10]Natural History Museum, University of Oslo, NO-0318 Oslo, Norway

Genomic comparisons provide evidence for ancient genome-wide duplications in a diverse array of animals and plants. We developed a birth–death model to identify evidence for genome duplication in EST data, and applied a mixture model to estimate the age distribution of paralogous pairs identified in EST sets for species representing the basal-most extant flowering plant lineages. We found evidence for episodes of ancient genome-wide duplications in the basal angiosperm lineages including *Nuphar advena* (yellow water lily: Nymphaeaceae) and the magnoliids *Persea americana* (avocado: Lauraceae), *Liriodendron tulipifera* (tulip poplar: Magnoliaceae), and *Saruma henryi* (Aristolochiaceae). In addition, we detected independent genome duplications in the basal eudicot *Eschscholzia californica* (California poppy: Papaveraceae) and the basal monocot *Acorus americanus* (Acoraceae), both of which were distinct from duplications documented for ancestral grass (Poaceae) and core eudicot lineages. Among gymnosperms, we found equivocal evidence for ancient polyploidy in *Welwitschia mirabilis* (Gnetales) and no evidence for polyploidy in pine, although gymnosperms generally have much larger genomes than the angiosperms investigated. Cross-species sequence divergence estimates suggest that synonymous substitution rates in the basal angiosperms are less than half those previously reported for core eudicots and members of Poaceae. These lower substitution rates permit inference of older duplication events. We hypothesize that evidence of an ancient duplication observed in the *Nuphar* data may represent a genome duplication in the common ancestor of all or most extant angiosperms, except *Amborella*.

[Supplemental material is available online at www.genome.org.]

Gene duplication has long been recognized to be a major force in evolution (Ohno 1970). Genome doubling (polyploidy) has had a profound influence on the evolutionary history of extant lineages. Ohno proposed that whole-genome duplications occurred in the early history of all vertebrates (Ohno 1970). While the hypothesis of whole-genome duplication in the earliest vertebrates has been somewhat controversial (Hughes 1999; Friedman and Hughes 2001; Makalowski 2001; Hughes and Friedman 2003), ancient polyploidy is supported by genetic and genomic investigations of individual gene families as well as large syntenic chromosomal segments (Abi-Rached et al. 2002; Gu et al. 2002; McLysaght et al. 2002; Dehal and Boore 2005). The importance of genome duplication in the evolution of amphibians (Bogart 1979) and the yeast *Saccharomyces cerevisiae* has been more widely accepted (Wolfe and Shields 1997; Friedman and Hughes 2001; Kellis et al. 2004).

Polyploidy is common in many plant lineages, particularly angiosperms (Stebbins 1950; Grant 1981; Soltis and Soltis 1999). The angiosperms in particular have been the subject of considerable speculation regarding the frequency of polyploidy. Classic studies estimated that 30%–50% of angiosperms are polyploids (Müntzing 1936; Darlington 1937; Stebbins 1950), and more recently most if not all extant angiosperms have been implicated as ancient polyploids (Grant 1963; Masterson 1994; Otto and Whitton 2000). Some of these inferences were based on comparisons of nuclear DNA content (C-value) or sequenced genome size, across a broad spectrum of species. However, the rapid reduction of duplicate genes immediately after polyploidization can drastically shrink genome size and gene content (Ohno 1970; deWet 1979; Liu and Wendel 2003). Despite the small size of the *Arabidopsis thaliana* genome (157 Mb) (Bennett et al. 2003), recent investigations have revealed two or more rounds of ancient genome duplications (Vision et al. 2000; Simillion et al. 2002; Bowers et al. 2003). Analysis of the rice genome also suggested ancient polyploidy in the early history of the grass family (Poaceae) (Paterson et al. 2004b; Yu et al. 2005). It now appears that perhaps all major lineages of eukaryotic genomes possess considerable numbers of duplicate genes that may have resulted from genome duplications (Ohno 1970; Lynch and Conery 2000).

Whole-genome duplication, tandem gene duplication, and segmental duplication all generate paralogous gene pairs. For species with complete genome sequences, such as *Arabidopsis*, rice, and now *Populus*, it is possible to differentiate whole-

[11]Corresponding author.
E-mail cwd3@psu.edu; fax (814) 865-9131.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.4825606.

16:738–749 ©2006 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/06; www.genome.org

genome duplications from segmental and tandem duplications by mapping chromosomal locations of duplicate genes or blocks of genes (Simillion et al. 2002; Blanc et al. 2003; Bowers et al. 2003; Cannon et al. 2004; Paterson et al. 2004b). Lynch and Conery (2000) proposed a genomic-scale approach to estimate the age of gene duplication events and the fate of resulting paralogous gene pairs by evaluating the frequency distribution of per-site synonymous divergence levels (K_s) for pairs of duplicate genes. After gene duplication, some paralogs will be silenced and eventually be eliminated, while many of the preserved paralogs may be subject to changes in DNA sequence or gene expression, leading to sub- or neofunctionalization (Force et al. 1999; Adams et al. 2003; Wang et al. 2004b).

Synonymous substitutions are largely immune to the strong selective pressures that greatly impact the rate of protein divergence (Li and Grauer 1991; Lynch and Conery 2000), and when corrected for multiple substitutions that occur in highly diverged sequences, these nearly neutral substitutions in protein-coding regions can be used as a proxy for the amount of time that has passed since gene duplication. A genome-wide duplication event simultaneously creates thousands of paralogous pairs. Evidence of past genome duplications can be seen as peaks in the distribution of K_s values for sampled paralogous pairs (Lynch and Conery 2000; Blanc and Wolfe 2004; Schlueter et al. 2004). This method does not depend on genomic positional information, and can be applied to any species for which there are moderately large EST sets. Identification of duplicated blocks of genes in genome sequences, however, provides much stronger evidence of ancient polyploidy, although average K_s values (or K_a) (Vision et al. 2000) can still be used to date the origin of duplicated blocks. Using the large number of DNA sequences generated by EST and genome sequencing projects, Blanc and Wolfe (2004) investigated 14 model plant species (mostly crop species with known recent polyploid history) and found spikes in the distributions of older paralogous pairs (with higher K_s values) in nine species. Schlueter et al. (2004) advanced the analysis of K_s distributions by applying a finite mixture model (McLachlan et al. 1999) to sets of paralogous pairs identified in large EST databases for eight major crop species, including soybean, *Medicago*, tomato, potato, maize, *Sorghum*, rice, and barley, and inferred multiple independent genome duplications in Fabaceae, Solanaceae, and Poaceae over the last 14–60 million years. In general, this method is only suitable for duplicated genes with similar codon usage, because K_s is affected by codon usage bias (Bierne and Eyre-Walker 2003; Wang et al. 2004a).

All of the plants previously investigated using K_s distributions (Blanc and Wolfe 2004; Schlueter et al. 2004) belong to either derived monocot (a single family, the Poaceae) or eudicot lineages. Most species examined were either crop species or close relatives, where a predisposition to polyploidy might have increased the chances of having traits important for domestication and agriculture (but see Hilu 1993). Until recently, there has been very little sequence data available for phylogenetically pivotal taxa representing the basal lineages of the eudicots, monocots, or all angiosperms, and the genome histories of these lineages are therefore poorly understood. Here and throughout this paper we use the term "basal" when referring to a lineage that is sister to a larger clade containing all other members of a particular group. An understanding of ancient genome duplication in the basalmost angiosperm lineages is especially important in understanding the role of polyploidy in the origin and early diversification of flowering plants (e.g., Buzgo et al. 2005; De Bodt et al. 2005; Zahn et al. 2005a,b). We use sets of 9000–10,000 ESTs generated for species representing basal angiosperms and basal eudicot lineages (Albert et al. 2005) to assess the frequency of ancient genome duplications across all major extant angiosperm lineages (Table 1) and to evaluate whether these data can elucidate the timing of ancient genome duplication events in early angiosperm history.

To facilitate the interpretation of K_s distributions, we have modeled the gene birth-and-death process both with and without genome-wide duplication events. Our model provides a predicted age distribution for any sample of duplicate genes while accounting for empirical estimation errors in K_s. The model was used to generate predicted K_s distributions for sets of paralogous pairs under the null hypothesis that the given gene births and deaths occurred at constant rates. Null distributions were modeled using parameter values and error corrections estimated for each data set (see Methods). When the null hypothesis of a constant birth-and-death process was rejected, the log-transformed K_s distribution for each taxon was analyzed using a mixture model to identify subpopulations of paralogous pairs generated through one or more large-scale duplication events (McLachlan et al. 1999; Schlueter et al. 2004). Our results provide evidence of

Table 1. Genome sizes and base chromosome numbers for the angiosperm and gymnosperm species in this study

Scientific name	Common name	Family	Group	Genome size (Mb/1C)	Chromosome number (2*n*)	Source
Arabidopsis thaliana	Thale cress	Brassicaceae	Rosid	157	10	RBG, Kew
Glycine max	Soybean	Fabaceae	Rosid	1103	40	RBG, Kew
Solanum lycopersicum	Tomato	Solanaceae	Asterid	1005	24	RBG, Kew
Eschscholzia californica	California poppy	Papaveraceae	Ranunculales	502	12	This study
Acorus americanus	Sweet flag	Acoraceae	Monocot	392	24	This study
Liriodendron tulipifera	Yellow-poplar	Magnoliaceae	Magnoliid	1710	38	This study
Persea americana	Avocado	Lauraceae	Magnoliid	907	24	RBG, Kew
Saruma henryi		Aristolochiaceae	Magnoliid	3014	52	This study
Nuphar advena	Yellow water lily	Nymphaeaceae	Basalmost angiosperm	2772	34	This study
Amborella trichopoda		Amborellaceae	Basalmost angiosperm	870	26	RBG, Kew
Pinus taeda	Loblolly pine	Pinaceae	Gymnosperm	21,658	24	RBG, Kew
Pinus pinaster	Pine	Pinaceae	Gymnosperm	23,863	24	RBG, Kew
Welwitschia mirabilis		Welwitschiaceae	Gymnosperm	7056	42	RBG, Kew

The relationships among the organisms and the major lineages are indicated in Figure 6. The sources for genome size data are the Royal Botanic Gardens, Kew Plant C-value database (RBG, Kew; http://www.rbgkew.org.uk/cval/homepage.html) and this study—the DNA content determined by flow cytometry as described in Wang et al. (2005).

ancient polyploidy throughout the major angiosperm lineages, and support the possibility that a genome-scale duplication event occurred prior to the rapid diversification of flowering plants (Darwin 1903).

Results

Model parameters and their influence on the observed age distribution of paralogs

To add statistical rigor to the interpretation of K_s distributions for paralogous pairs, we modeled the expected age and K_s distributions under a constant-rate birth–death model (see Methods). Whereas recent studies have shown that evidence of paleopolyploidy is often (but not always) discernible in K_s plots for paralogous pairs (Blanc and Wolfe 2004; Schlueter et al. 2004; Maere et al. 2005), the accumulation of single gene duplications, variation in the rates of gene death, and error in K_s estimates have not been studied quantitatively. We model the rate of gene death, the time since gene (or genome) duplication, and the error in K_s estimates in analyses of paralogous pairs. Our null model assumes gene birth and death are independent events, each with a constant rate over time. Under this model, the expected age distribution for paralogous pairs is a declining exponential with a decay parameter corresponding to the rate of gene death. K_s distributions derived from simulations under this model are influenced by the random nature of nucleotide substitution and the error in K_s estimation. To formally test for deviation from a constant-rates' model using empirical data, we generate a null distribution for the frequency of K_s values using parameters estimated from the data for the rate of gene death and the error in K_s estimation.

Our model was also used to simulate K_s distributions for paralogous pairs arising from a mixture of single gene duplications and ancient polyploidy events. Empirical estimates of variation in K_s were based on analyses of *A. thaliana* paralogous pairs. Figure 1 shows K_s distributions for data simulated with different rates of gene death and different times since a genome duplication event. These K_s distributions contain two components; the first one is always a declining exponential distribution corresponding to "background" single gene duplications, and the second component represents paralogous pairs arising from a polyploidy event. Very recent genome duplications may be obscured by background gene duplications because the modal K_s values do not appear as distinct peaks. Conversely, increases in the number of gene deaths and variance in K_s with time render older genome duplications less detectable than younger events, with no significant duplication signal recovered for events with an expected K_s of 1.5 (Fig. 1C,F,I). High gene death rates also eroded the impact of genome duplications on K_s distributions (Fig. 1G,H,I). These results corroborate previous evidence that ancient genome duplication events are not always detectable in analyses of K_s distributions (Blanc and Wolfe 2004; Paterson et al. 2004a).

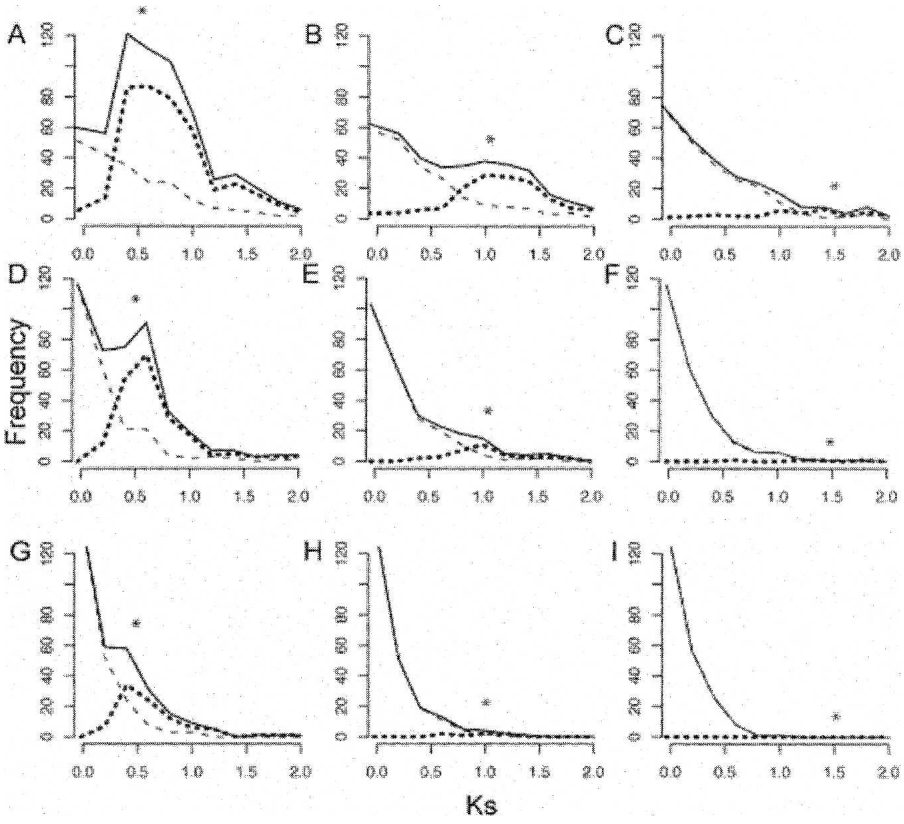

Figure 1. Effect of gene death rate and time of genome duplication on the K_s distribution for paralogs. A single genome duplication was simulated, where time since duplication (corresponding to K_s = 0.5 in *A,D,* and *G*; 1.0 in *B,E,* and *H*; or 1.5 in *C,F,* and *I*) is indicated by a star. The death rate of duplicate pairs (δ) increases from the *top* row to the *bottom* row (δ = 0.67 for *A,B,C,* as estimated from *Arabidopsis* data; 1.34 for *D,E,F*; and 2.68 for *G,H,I*). In each graph, the observed frequency of paralogs from background gene duplication is plotted with a dashed line, while the distribution deriving from genome duplication is plotted with a dotted line. The K_s distribution of all paralogs is drawn with a solid line.

Evidence of genome duplications in diverse lineages of flowering plants

Model validation: Duplications detected in eudicots

EST sets from *A. thaliana*, *Glycine max* (soybean), and *Solanum lycopersicum* (tomato) were used to validate our test of the constant-birth–death-rate model. The genome duplication histories for these species have been elucidated in several previous analyses (Shoemaker et al. 1996; Blanc et al. 2000; Grant et al. 2000; Ku et al. 2000; Vision et al. 2000; Bowers et al. 2003). To make these analyses comparable to analyses of the other EST sets in this study, we randomly sampled sets of 6000 unigenes, or ~10,000 ESTs, from a much larger set of available ESTs for each of these taxa (see Methods).

To determine whether inference of genome-wide duplication events depends on the method of synonymous substitution estimation, we compared four methods of K_s estimation, including the original Nei-Gojobori (NG) method (Nei and Gojobori 1986), the modified Nei-Gojobori (modified NG) method (Zhang et al. 1998), the Goldman and

Yang maximum likelihood (ML) method (Goldman and Yang 1994), and the YN00 (YN) method (Yang and Nielsen 2000). Results were similar across all K_s estimation procedures in analyses of the *Arabidopsis* data set (Fig. 2A). Analyses of replicate subsamples from the random *Arabidopsis* unigenes gave very similar results to analyses of all paralogous pairs (Fig. 2B; Lynch and Conery 2000; Blanc et al. 2003; Maere et al. 2005), suggesting that 6000 unigenes are sufficient for estimating K_s distributions for the other species in this study (Table 2).

We estimated the rate parameter for *Arabidopsis* data ($\delta = 0.67$) assuming a constant-birth–death model (the null model) and tested the expected distribution against the observed distribution using a χ^2 test (Fig. 2C). The null model was rejected ($P \ll 0.0001$), and the quantile–quantile plot showed obvious deviation from the expected distribution of K_s values (bootstrapped Kolmogorov-Smirnov test, $P \ll 0.0001$) (Fig. 2D). Next, we applied the mixture model to estimate the median age (in K_s equivalent units) of duplicate genes from recent or older duplication events (Table 3). This analysis, using ML distances, identified two significant components, a background component with median $K_s = 0.2889$, and a prominent second component including 79% of the paralogous pairs with a median $K_s = 0.7510$ that corresponded to the polyploidy peak detected by Blanc and Wolfe (2004). Similar results were obtained when the YN, NG, and modified NG K_s estimates were used, thus only ML distance estimates are reported for all other analyses since they are typically less biased and have lower error rates, especially for more divergent sequences (Yang and Nielsen 2000). We obtained similar results to those reported in previous studies (Blanc and Wolfe 2004; Schlueter et al. 2004), with much smaller subsamples of ESTs (Fig. 2B).

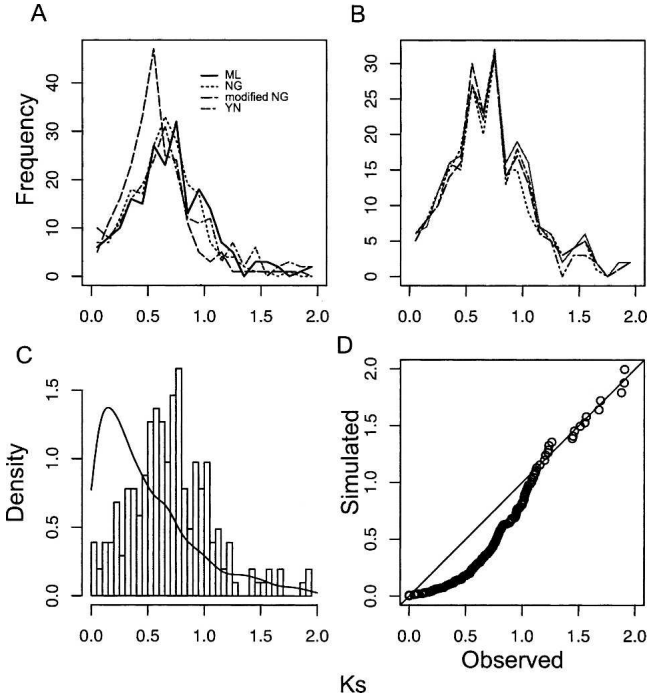

Figure 2. K_s distribution from a sample of *Arabidopsis* unigenes and the diagnostic test according to the constant birth–death model (null model). (*A*) K_s estimates from four methods show strong agreement. (ML) Maximum likelihood method by Goldman and Yang; (NG) Nei-Gojobori method; (mNG) modified Nei-Gojobori method; (YN) Yang and Nielsen method. These sample sizes are comparable to the unigenes available for the species sequenced in this study. (*B*) K_s distributions for paralogs from four replicate unigene samples of 6000 sequences each. (*C*) The density plot of observed K_s distribution and simulated data based on the null model with parameter $\delta = 0.67$. (*D*) The Q-Q plot of observed versus expected K_s values shows the poor fit of the null hypothesis that gene birth and death rates are constant ($P \ll 0.0001$).

Table 2. Summary of EST data sets and paralogous pairs identified in this study

Scientific name	ESTs	Unigenes	Pairs with $K_s < 2$	Source
Arabidopsis thaliana		6000[a]	205	DbEST
Glycine max	10,046	6240	125	DbEST
Solanum lycopersicum	10,028	5303	143	DbEST
Eschscholzia californica	9079	5713	178	PGN
Acorus americanus	7484	4663	149	PGN
Liriodendron tulipifera	9531	6520	92	PGN
Persea americana	8735	6183	196	PGN
Saruma henryi	10,273	6293	184	PGN
Nuphar advena	8442	6205	138	PGN
Amborella trichopoda	8629	6099	69	PGN
Pinus taeda		6000[b]	276	PlantGDB
Pinus pinaster		6000[c]	259	PlantGDB
Welwitschia mirabilis	9776	6048	157	PGN

[a]Sampled from 6369 unigenes.
[b]Sampled from 52,527 unigenes.
[c]Sampled from 8076 unigenes.

We next analyzed public EST sets from selected libraries of soybean and tomato. Soybean ESTs were sampled from flower, young seedling, root, and other vegetative tissue libraries. Mixture model analysis suggests that 71% of the paralogs were likely to have arisen from a large-scale duplication (Table 3), which appears as a significant peak in the K_s distribution with estimated median $K_s = 0.6705$ (Fig. 3A). This species is a relatively recent tetraploid (Shoemaker et al. 1996; Blanc and Wolfe 2004; Schlueter et al. 2004). Thus many of the duplicate pairs assigned to the first component in the mixture model are likely derived from polyploidization rather than background single gene duplications.

The results for tomato also suggest large-scale duplications, which account for >90% of paralogs. Moreover, the distributions for paralogous gene pairs sampled from two tissue sources (floral vs. nonfloral organs) were similar (Fig. 3B,C) and in agreement with previous analyses based on all duplicate gene pairs in this species (median $K_s = 0.277$ and 0.632) (Schlueter et al. 2004). Together, our tests found strong signals of deviation from the null model, and as expected, mixture model analyses suggested ancient polyploidy events in *Arabidopsis*, *Glycine*, and *Solanum*. Taken together, these results suggest that unbiased K_s distributions can be obtained from as few as 6000 unigenes sampled from complex cDNA libraries derived from developing floral organs.

Ancient polyploidy in a basal eudicot

Eschscholzia californica (California poppy, Papaveraceae) is a member of Ranunculales, the sister lineage to all other eudicots (Soltis et al. 2000; Zanis et al. 2002; Borsch et al. 2003). Analysis of the K_s distribution of 178 pairs of *Eschscholzia* paralogs rejected the constant birth and death model ($P \ll 0.0001$), and two components in the distribution were identified by the mixture model. The second component dominated the distribution, with 89% of the duplicate pairs (Fig. 3D), providing the first strong

Table 3. Mixture model estimates for K_s distributions in each species

Scientific name	*n*	*P*	lnL	BIC	Median	Variance	Proportion
Arabidopsis thaliana	202	2	−162.498	351.54	0.2889	0.0473	0.21
					0.751	0.0777	0.79
Glycine max	123	2	−147.358	318.78	0.1873	0.0398	0.29
					0.6705	0.1066	0.71
Solanum lycopersicum (floral)	139	2	−118.607	261.89	0.0643	0.0066	0.09
					0.7894	0.1021	0.91
Solanum lycopersicum (nonfloral)	119	2	−122.933	269.76	0.1857	0.0547	0.15
					0.7885	0.1425	0.85
Eschscholzia californica	178	2	−161.652	349.21	0.0871	0.0043	0.11
					0.7098	0.087	0.89
Acorus americana	139	3	−103.568	246.61	0.0118	0.001	0.01
					0.455	0.0046	0.33
					0.5813	0.1309	0.65
Liriodendron tulipifera	87	2	−94.046	210.42	0.1005	0.0121	0.14
					0.7616	0.1328	0.86
Persea americanus	186	2	−196.998	420.12	0.0234	0.0004	0.07
					0.6464	0.1197	0.93
Saruma henryi	146	2	−162.789	350.5	0.0913	0.0168	0.2
					0.7927	0.1066	0.8
Nuphar advena	134	3	−159.416	358.02	0.1746	0.0461	0.37
					0.4291	0.0202	0.56
					1.3273	0.0084	0.07
Amborella trichopoda	49	1	−80.676	169.14	0.2698	0.1147	1
Pinus taeda	227	1	−405.77	822.39	0.0839	0.0147	1
Pinus pinaster	240	1	−373.135	757.23	0.2499	0.0819	1
Welwitschia mirabilis	132	2	−181.128	386.67	0.1139	0.0271	0.35
					0.9519	0.1374	0.65

Initial tests against the null model (no genome duplication) were conducted, then a mixture analysis was applied to each species. The final mixture model was selected according to the Bayesian Information Criterion (BIC) and restriction on the mean/variance structure for K_s (see Methods). (*n*) Sample size; *P*, number of mixture components, −lnL, log likelihood for the mixture model. For each mixture model, the proportions for each component (subpopulation) sum to 1.

evidence of probable ancient genome duplication in a basal eudicot. Phylogenetic analyses of duplicated *AGAMOUS* and *AP3* homologs (Kramer et al. 1998; Kramer and Irish 1999; Zahn et al. 2005a) have suggested that this duplication event occurred after the split between Ranunculales and core eudicots. Thus, the genome-wide duplication evident in the *Eschscholzia* paralogous pairs was probably independent of the genome duplications that have been inferred from analyses of the *Arabidopsis* genome (Vision et al. 2000; Bowers et al. 2003; Maere et al. 2005).

Basal monocot

Acorus americanus (Acoraceae, Acorales) represents the sister lineage to all other monocots (Duvall et al. 1993; Soltis et al. 2000, 2002; Zanis et al. 2002; Borsch et al. 2003; Hilu et al. 2003). Three components were identified in the paralogous pairs by the mixture model approach. The second component, accounting for 33% of all duplicates, was shown as a sharp peak in the K_s distribution, while the third component, containing 65% of the duplicates, appeared as a broader peak (Fig. 3E). Based on the distinct modes observed in the raw K_s distribution, we hypothesize that the second and third components estimated in the mixture model represent two distinct large-scale duplication events. This hypothesis will be tested in future phylogenetic analyses of well-sampled gene families.

Magnoliids

Both shared and lineage-specific genome duplications were inferred from analyses of unigenes from three magnoliid species: *Liriodendron tulipifera* (Magnoliaceae, Magnoliales), *Persea americana* (Lauraceae, Laurales), and *Saruma henryi* (Aristolochiaceae, Piperales). A total of 92 paralogous pairs was detected in the *Liriodendron* unigene set. The constant-birth–death model was rejected ($P < 0.001$), and a mixture of two components was identified in the K_s distribution, with the second component being dominant (Fig. 4A). The null birth–death model was also rejected in the *P. americana* (avocado) analysis ($P \ll 0.0001$) with 196 paralogous gene pairs. The optimal mixture model also included two components very similar to those seen for *Liriodendron* (Fig. 4B; Table 3).

To determine whether the duplication events inferred from the K_s distributions of *Liriodendron* and *Persea* represented events in a common ancestor, we first computed the median K_s of putatively orthologous gene pairs (408 pairs identified as reciprocal best hits in BLAST searches) and compared the median K_s for orthologs with K_s values for paralogous pairs within each species. The K_s distribution of putative ortholog pairs showed a single major component (median = 0.8057, variance = 0.0858) (Fig. 4F), inferred to be slightly older than the probable genome duplication observed in *Persea* (median = 0.6464, variance = 0.1197; $P < 0.0001$, Wilcoxon test). The timing of the duplication event inferred from the *Liriodendron* K_s distribution (median = 0.7616, variance = 0.1328) relative to the divergence of the *Persea* and *Liriodendron* lineages was ambiguous ($P = 0.35$), and direct comparison of the *Persea* and *Liriodendron* K_s distributions may have been confounded by unequal substitution rates.

To account for possible variation in synonymous substitution rates between the *Persea* and *Liriodendron* lineages, we aligned putatively orthologous genes from *Liriodendron*, *Persea*, and *Saruma* and estimated K_s values for each lineage on a phy-

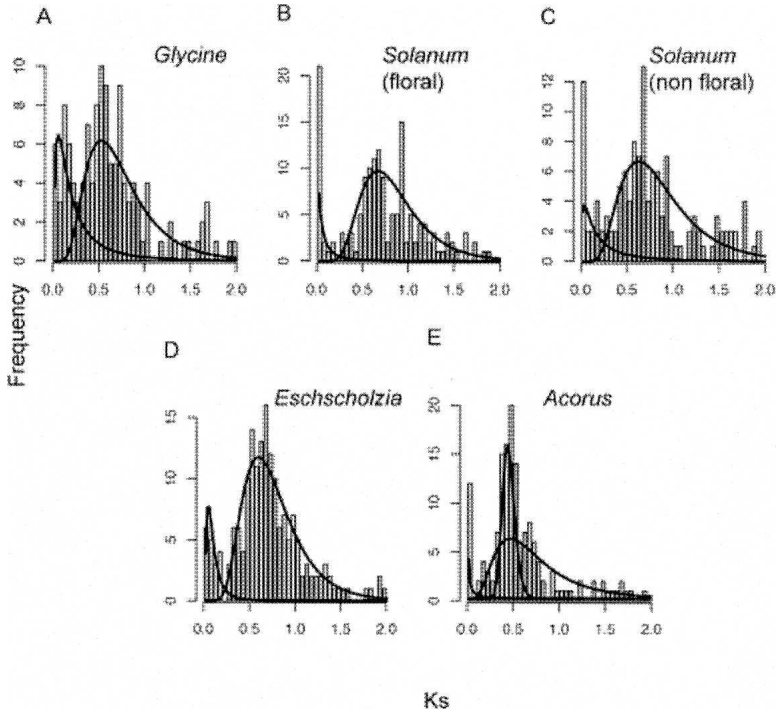

Figure 3. K_s distributions of paralogs in selected angiosperm species, with fitted densities from mixture model analysis, suggest paleopolyploidy in eudicots and monocots. Each fitted line indicates a subpopulation in the mixture. The first (*leftmost*) component corresponds to paralogs from background gene duplications; other peaks indicate estimated median K_s for ancient duplications. (*A*) *Glycine max* (soybean). (*B*,*C*) *Solanum lycopersicum* (tomato), data from floral tissue (*B*) and nonfloral tissue (*C*). (*D*) A basal eudicot, *Eschscholzia californica* (California poppy). (*E*) A basal monocot, *Acorus americanus*.

logeny. We examined 19 putative orthologous gene sets in the three species with alignments of at least 400 bp for all taxa (see Methods; Supplemental Table S2) and found that K_s on the lineage leading to *Liriodendron* was slower on average than the rate on the lineage leading to *Persea*. For example, in the tree for the orthologous set shown in Figure 4G, the branch length (in K_s units) for the branch leading to *Persea* is 1.31 times that leading to *Liriodendron*. The ratio of synonymous substitutions on the *Persea* branch relative to the *Liriodendron* branch ranged from 0.86 to 2.68, and the ratio was greater than one in 16 of 19 cases (Supplemental Table S2). When *Liriodendron* paralog K_s values were multiplied by the median branch-length ratio, 1.29, the peak in the scaled *Liriodendron* K_s distributions matched an older, but nonsignificant peak in the *Persea* K_s distribution (Fig. 4E). Taken together, these analyses suggest that the prominent peak in the *Liriodendron* K_s distribution (median = 0.82) represents a duplication event in the common ancestral genome of Magnoliales and Laurales that had not been identified as a distinct component in the mixture model for the *Persea* K_s distribution. In line with the comparison of K_s values for *Persea* paralogs and putative *Liriodendron–Persea* orthologs, we interpret the dominant peak in the *Persea* K_s distribution to represent a genome-scale duplication event that occurred after the divergence of Magnoliales and Laurales. This hypothesis needs to be tested with additional data.

Saruma henryi is a member of Piperales, which (with Canellales) is sister to the Magnoliales/Laurales clade (Soltis et al. 2000, 2002; Zanis et al. 2002; Borsch et al. 2003). The K_s distribution of *Saruma* paralogs showed a distinct peak with median K_s = 0.7927 (Fig. 4C; Table 3). This is lower than the median K_s for 202 *Saruma–Liriodendron* ortholog pairs (0.9555, $P = 0.0001$) and the median K_s for 254 putative *Saruma–Persea* ortholog pairs (1.0121, $P < 0.0001$) (Fig. 4F). We therefore surmise that the peak in the K_s distribution of *Saruma* paralogous pairs represents a large-scale duplication in Piperales after divergence from the Magnoliales and Laurales lineages.

Basal–most angiosperms

Amborella trichopoda (Amborellaceae) and the water lilies (Nymphaeales) are either successive sister lineages to all other extant angiosperms, or together form a clade that is sister to the rest of the angiosperms (Zanis et al. 2002; Stefanovic et al. 2004; Leebens-Mack et al. 2005). The K_s distribution for a total of 69 *Amborella* paralogous pairs appeared to follow an exponential distribution, but the uniform birth–death process was rejected ($P < 0.01$; Figure 5A). However, the mixture model analysis identified only one component containing all of the gene pairs (Table 3). Nymphaeales are represented by *Nuphar advena*. A total of 138 paralogous pairs was identified, and the resulting K_s distribution did not fit the constant birth–death model ($P < 0.01$). Three mixture components were estimated from the K_s distribution. The second component, accounting for 56% of the paralogous pairs, provided strong evidence for ancient polyploidy in the history of the *Nuphar* genome (Fig. 5B). The third component, with a median K_s of 1.3273, may represent the oldest genome duplication to be detected in analyses of angiosperm K_s distributions. The median K_s for the third component was not distinguishable from the median K_s value for putative *Amborella–Nuphar* orthologs (Fig. 5C) (median K_s[orthologs] = 1.24, variance = 0.1918, based on 113 putatively orthologous sequence pairs; $P = 0.05$, two-sample *t*-test on the log K_s[orthologs] and log K_s[third component of *Nuphar* paralogs]). Therefore, the third component in the *Nuphar* K_s distribution may correspond to a polyploidy event that occurred at approximately the time of divergence between the *Amborella* and *Nuphar* lineages (see Discussion).

Gymnosperms

We obtained 52,527 unigenes for *Pinus taeda* (loblolly pine) from PlantGDB (Dong et al. 2004), and a random sample of 6000 unigenes was drawn to match the sample size for other species we investigated. The K_s distribution showed a clear monotonous decay of paralogs with increasing age and no detectable sign of genome duplication in recent history ($P = 0.16$) (Fig. 5D). The frequency distribution for all paralogous pairs was essentially identical. The analysis of *Pinus pinaster* yielded a similar exponential distribution (Table 3).

The constant-birth–death model was rejected for *Welwitschia* ($P < 0.01$), and a mixture analysis of the K_s distribution

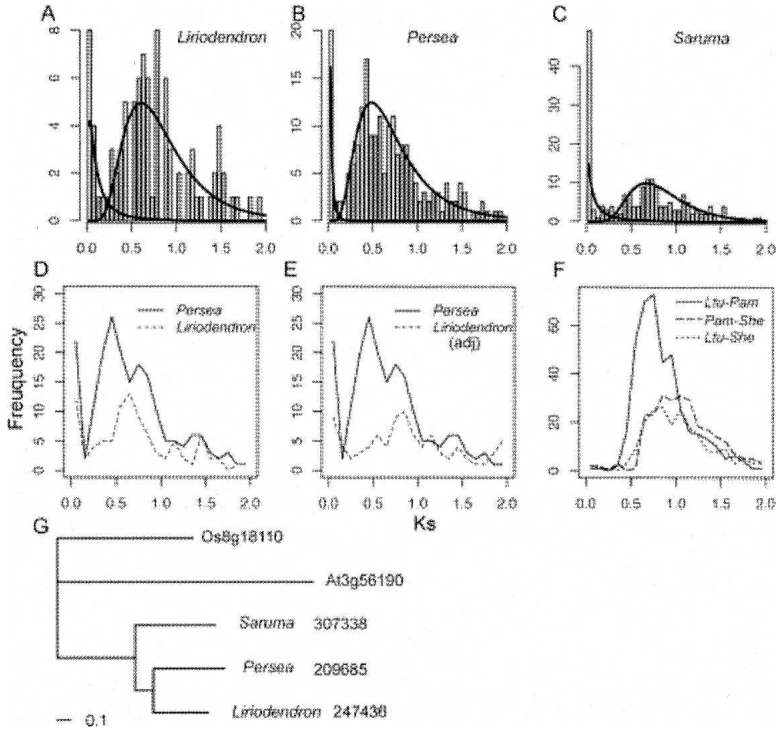

Figure 4. K_s distributions of paralogs and orthologs among magnoliids suggest independent duplications and possibly shared genome duplication events in Laurales (*Persea*) and Magnoliales (*Liriodendron*). (*A*,*B*,*C*) The K_s distributions for (*A*) *Liriodendron*, (*B*) *Persea*, and (*C*) *Saruma*, with fitted lines based on the mixture model analysis. (*D*) The K_s distribution for *Liriodendron* and *Persea*, without scaling for rate differences between lineages. (*E*) K_s distribution for paralogs in *Liriodendron* after rate calibration (adj = adjusted), compared with that of *Persea*, suggesting recent independent duplication and older shared genome-scale duplications. (*F*) K_s distribution for orthologs of two magnoliid species. (*Ltu*) *Liriodendron*; (*Pam*) *Persea*; (*She*) *Saruma*. (*G*) Phylogeny of one representative orthologous gene set used for relative rate estimates. The branch lengths show the estimated relative rates of synonymous evolution in respective species.

identified two components (Fig. 5E). The second component, corresponding to the heavy right-hand tail of the distribution, may represent one or more ancient duplication events, or a reduced rate of gene death for older duplicates.

Discussion

In this paper, we introduce a model-based statistical test that accounts for estimation error in K_s values in terms of deviation from a constant rate of gene birth and death. This represents a refinement of previous studies using K_s distributions, which have yielded significant insights into genome duplications (Force et al. 1999; Lynch and Conery 2000; Blanc and Wolfe 2004; Schlueter et al. 2004). The birth–death model developed here for duplicated genes is a natural extension of stochastic birth-and-death models that have been widely used in population and phylogenetic approaches to studies of gene family evolution (Karev et al. 2004). Simulations based on this model have allowed us to investigate how specified death rates and duplication times result in K_s distributions with (or without) secondary peaks or heavy tails (e.g., Fig. 1). The model can be extended to incorporate variable rates of gene birth or death over time, and in the extreme, an instant burst of gene birth corresponding to a whole-genome duplication. Although our results could not exclude partial and segmental duplication events, the birth–death model was validated with genomes with known duplication histories where detection of whole–genome events was expected.

We found that three major factors influence the frequency and observed divergence of paralogous pairs arising from genome-wide duplications. The time since the duplication event, the rate of gene death, and the background rate of gene birth all influence observed K_s distributions. Very recent genome duplication events are associated with K_s values for resulting paralogous pairs that are indistinguishable from those of background single-gene duplications using EST data. For example, polyploidy is not clearly evident in the K_s distribution for hexaploid wheat because there has been little divergence among the parental or homeologous gene copies, and the range of divergence for allelic variants was not distinct from that of paralogs arising from recent gene duplications (Blanc and Wolfe 2004). At the same time, evidence of very ancient genome duplications is eroded as synonymous substitutions reach saturation and variance in K_s increases. This may be evident in K_s plots for wheat, maize, rice, and barley, for which evidence for a genome duplication event some 50–60 million years ago (Mya) in the common ancestor of all major grain lineages has been obscured (Blanc and Wolfe 2004; Paterson et al. 2004a). Detection of very old duplication events in K_s distributions is especially difficult in species with high synonymous substitution rates. Conversely, evidence for the oldest detectable genome-wide duplications will be found in K_s distributions for species with the slowest substitution rates (see below).

Concurrent expansion of a few gene families could lead to moderate deviations from our null model. This is especially true if ancient duplication events are overrepresented in sets of sampled paralogous pairs, or if major adaptive radiations of individual gene families preceded or accompanied the diversifications of the organismal lineages under study. In this study, we avoided over-counting of ancient gene duplications by constraining genes to be included in only one paralogous pair. Our analysis of duplicated *Arabidopsis* genes verified that this approach produced K_s distributions similar to those of previous studies that implemented more elaborate corrections for gene family expansions (Maere et al. 2005). Moreover, sampled paralogous genes were not particularly biased toward large gene families. Whereas most sampled duplicate genes belonged to the housekeeping functional categories, such as protein synthesis, proteolysis, and energy metabolism (Supplemental Table S1), none of the duplicate gene sets was dominated by a single gene family. Several transcription factor families were also identified in our paralog pairs, but again, no family accounted for more than a few percent of the duplicate gene pairs.

Our results for *Persea* (Lauraceae) and *Liriodendron* (Magnoliaceae) corroborate previous evidence of ancient polyploidy

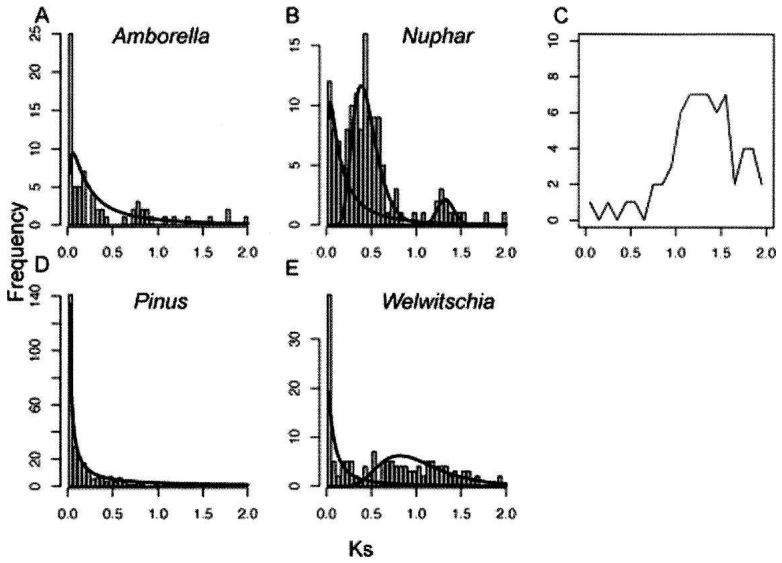

Figure 5. K_s distributions suggest possible genome duplications in basal angiosperms, and no evidence for genome duplication events in *Amborella* and some gymnosperm species. (*A*) K_s distribution in *Amborella*, a basal-most angiosperm. No significant large-scale duplication is detected. (*B*) Three distinct components in the K_s distribution for *Nuphar*, also a basal-most angiosperm, suggest at least two large-scale genome duplications. (*C*) K_s distribution for putative orthologs between *Amborella* and *Nuphar*. (*D*) *Pinus taeda* (loblolly pine) paralogous pairs follow the null model (see Methods). (*E*) K_s distribution for paralogs in a gymnosperm, *Welwitschia*.

from isozyme studies (Soltis and Soltis 1990). Soltis and Soltis (1990) found that 25%–29% of the loci investigated were duplicated in both families, and hence could have arisen via polyploidy. All members of Magnoliaceae examined shared the same isozyme duplications (PGI, TPI, 6PGD), while the species of Lauraceae shared a similar suite of isozyme duplications (PGM, TPI, 6PGD, GDH). These were interpreted as evidence for paleopolyploid events occurring very early in the evolutionary history of Magnoliaceae and Lauraceae. The *Persea* and *Liriodendron* paralogous genes suggest polyploidy in a common ancestor at least 100 Mya (Bell et al. 2005) followed by a second round of polyploidy in the *Persea* lineage (Fig. 4E), but this hypothesis must be tested with analyses of additional gene family phylogenies. If this scenario is correct, the duplicated isozyme loci observed in the Magnoliaceae and Lauraceae may have arisen from a polyploidy event that predated the separation of the two families (cf. Brysting and Borgen 2000).

Over time, nucleotide substitutions can become saturated, and therefore lineages with slow synonymous substitution rates will provide a deeper view into genome history relative to lineages with faster substitution rates. It is estimated that the synonymous substitution rate in palm (2.61×10^{-9} synonymous substitutions/per year) (Gaut et al. 1996) is only about half that reported for grasses, eudicots (Lynch and Conery 2000), and grass–eudicot comparisons (Wolfe et al. 1987). We infer a similarly slow substitution rate for other basal angiosperms based on the Magnoliales–Laurales divergence as a calibration point. We estimated a synonymous site divergence of $K_s = 0.7$ for *Liriodendron* and *Persea* ortholog pairs (Fig. 4F). Using a divergence date estimate of ~116 Mya for the Magnoliales–Laurales split (Bell et al. 2005), we estimate an average synonymous substitution rate of 3.02×10^{-9} synonymous substitutions/year. The low substitution rate in *Liriodendron* and *Persea* may be explained in part by their longer generation times (these lineages are trees and shrubs) relative to model eudicot and grass species.

We found that the median for the oldest component in the *Nuphar* K_s distribution is close to the median K_s for putative *Amborella–Nuphar* orthologs (median $K_s = 1.24$) (Fig. 5C). This level of divergence is compatible with the synonymous divergence for the very early duplication in *Arabidopsis* (i.e., γ duplication) (Bowers et al. 2003; De Bodt et al. 2005; Maere et al. 2005). Direct dating of the early *Nuphar* peak based on the K_s data is challenging because of uncertainty in the branching relationships between *Amborella*, *Nuphar*, and the rest of the angiosperms, and the possibility of additional rate variation as was seen for magnoliids. We adopted two approaches to date the earliest event in *Nuphar*. First, using the median K_s *Amborella–Nuphar* ortholog divergence of 1.24 and a calibration range of 134–165 Mya (Leebens-Mack et al. 2005) gives a rate of $4.66–3.79 \times 10^{-9}$ substitutions per silent site per year. Therefore, $K_s = 1.33$ (the early *Nuphar* duplication event) would predict an age range between 143 and 173 Mya for the split between these two lineages. An alternative calculation, using the magnoliid calibration of 3.02×10^{-9} substitutions per silent site per year, leads to an estimate of 220 Mya for the divergence of lineages leading to *Amborella* and *Nuphar*.

This range of age estimates supports two alternative interpretations of the *Nuphar* and *Amborella* paralog K_s distributions. The third component in the *Nuphar* K_s distribution may represent polyploidy in a common ancestor of all angiosperms (Fig. 6), in agreement with recent analyses of MADS-box gene families (Kim et al. 2004; Buzgo et al. 2005; Zahn et al. 2005a). This scenario would require that evidence of ancient polyploidy has been sufficiently eroded as to be undetected in analyses of EST samples from *Amborella* and various other angiosperm species owing to gene death and/or saturation of synonymous substitutions as discussed above. For example, the nonsignificant peaks around $K_s = 1.5$ in the *Liriodendron* and *Persea* K_s distributions (Fig. 4A,B) may provide weak evidence of polyploidy early in angiosperm history. Alternatively, the earliest duplication peak detected in the *Nuphar* analysis may trace back to a genome duplication in the common ancestor of *Nuphar* and all extant angiosperm lineages other than *Amborella* (Fig. 6). This scenario would be consistent with the hypothesis that *Amborella* is sister to all other extant angiosperms (e.g., solid line on Fig. 6), and the extremely low proportion of duplicate genes found in the *Amborella* unigene set. This scenario also would narrow the timing of a genome duplication to ~10 Myr separating the branching points for *Amborella* and all other extant angiosperm lineages (Leebens-Mack et al. 2005). As discussed above, however, there have been instances where known genome duplication events have not been detected in K_s distributions (Fig. 1; Blanc and Wolfe 2004; Paterson et al. 2004b), thus lack of evidence for ancient polyploidy in the *Amborella* K_s distribution does not exclude the possibility of polyploidy in an ancestral genome. More

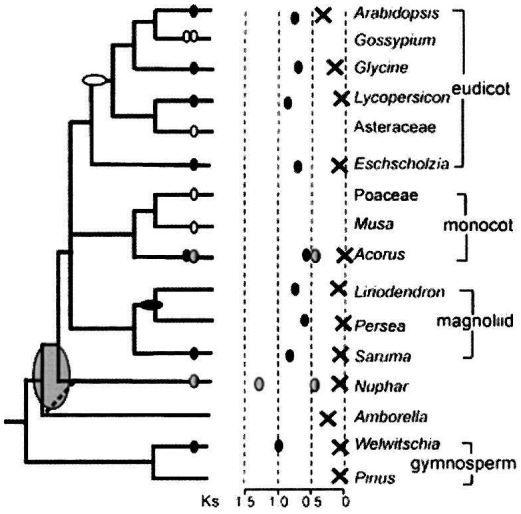

Figure 6. Phylogenetic summary of paleopolyploidy events estimated by the mixture model approach and their distribution among angiosperm and gymnosperm lineages. Scaled graph in center with Xs corresponding to median K_s of pairs from background gene duplications, while small ovals indicate the median K_s of possible concentrated duplications in the history of particular lineages. The phylogenetic tree at *left* shows the likely placement of detected genome-scale duplications. Uncertainty in phylogenetic timing of what may be a single duplication event at the base of the angiosperms is indicated with a wide oval that covers possible branching points compatible with the K_s evidence. Hollow ovals indicate duplications identified in previous studies using paralogous genes or genomic data from those lineages.

sequence data, and ultimately whole genome sequences, will be needed from *Amborella*, water lilies, and other early branching angiosperm species to select among these alternative scenarios for polyploid origins of angiosperms.

While genomic sequences have revealed evidence of polyploidy in Poaceae and core eudicots, the secondary peaks found in paralog K_s distributions for representatives of virtually all major angiosperm lineages support the notion that genome duplications are common in angiosperm history and gene birth and death are important processes in plant evolution (Lynch and Conery 2000). The evidence now supports the hypothesis proposed initially decades ago by Stebbins (1950) that angiosperms have experienced repeated rounds of polyploidization throughout their evolutionary history. Many questions follow: How many polyploidy events separate different plant lineages? What is the typical fate of genes generated through these duplication events? And perhaps most intriguingly, have polyploidy events been important engines of angiosperm diversification? Genome-scale sequencing of phylogenetically crucial angiosperm species would provide the data necessary to directly test whether the rapid diversification of flowering plants following their origin (Darwin 1903) was associated with one or more polyploidy events.

Methods

EST sequencing and assembly

EST sequences from floral cDNA libraries of seven species (*Amborella trichopoda*, yellow water lily [*Nuphar advena*], avocado [*Persea americana*], yellow-poplar [*Liriodendron tulipifera*], wild ginger [*Saruma henryi*]), sweet flag [*Acorus americanus*], and California poppy [*Eschscholzia californica*]) are available through the Plant Genome Network (http://www.pgn.cornell.edu). cDNA library construction, EST sequencing, and assembly were described previously (Albert et al. 2005).

Public EST sets from selected libraries for *Arabidopsis thaliana*, soybean (*Glycine max*, Williams 82), and tomato (*Solanum lycopersicum*, cultivar TA496) were downloaded from the GenBank dbEST section, trimmed using seqclean, and assembled using CAP3 with the percent identity parameter $P = 90$ and overlap length 40 bp. *A. thaliana* ESTs were from four libraries (root, flower, green silique, and 2- to 6-wk above-ground organs). To minimize the allelic variations in the EST sequence collection, the unigenes were mapped to the *Arabidopsis* genome, and redundant unigenes matching the same genomic locus were discarded. Only the sequences that matched the protein-coding regions were retained. From this screened unigene set, we drew replicate samples with 6000 unigenes in each sample. The sample size of 6000 *Arabidopsis* unigenes approximates the number of unigenes from new EST data sets we analyzed. To see if library sources influence estimates, we analyzed two samples of tomato ESTs, one from floral cDNA libraries and one from vegetative cDNA libraries. The soybean ESTs were sampled from cDNA libraries of flower, young seedling, root, and other vegetative organs. Unigenes for gymnosperms *Pinus taeda* and *Pinus pinaster* were downloaded from PlantGDB (Dong et al. 2004), which were built with public ESTs from all libraries. For each species, we sampled 6000 unigenes for K_s analysis.

K_s calculation for paralogs and orthologs

Paralogous pairs of sequences were identified from best reciprocal matches in all-by-all BLASTN searches. For data sets with trace files, we discarded bases with Phred (Ewing and Green 1998; Ewing et al. 1998) quality values lower than 20. Only sequence pairs with alignment lengths >300 bp were used for K_s calculations. Translated sequences of unigenes generated by ESTScan (Iseli et al. 1999) were aligned using MUSCLE v3.3 (Edgar 2004). Nucleotide sequences were then forced to fit the amino acid alignments. The K_s value for each sequence pair was calculated using the Goldman and Yang maximum likelihood method (Goldman and Yang 1994) implemented in codeml with the F3×4 model (Yang 1997). In order to assess whether the shape of K_s distributions was dependent on the estimation procedure, the Nei-Gojobori method, the modified Nei-Gojobori method, and the YN00 method (Yang and Nielsen 2000) were also applied on the *Arabidopsis* set. The K_s frequency in each interval size of 0.05 within the range [0, 2.0] was plotted.

The age distribution of paralogs under a constant birth–death model (the null model)

We modeled the birth and death of paralogs formed by gene duplications under a constant-rate birth–death model in order to test whether an observed frequency distribution of K_s values indicates deviation from this process. The duplicate genes are generated by a Poisson process at rate β, and the number of duplicate pairs decreases by age at an exponential rate δ. We can estimate the age distribution of surviving paralogs (survivors), total N, by considering the process as sampling gene birth over time [0, t], and decide if each birth was a survivor.

The distribution for the number of survivors of age t is

$$N(t) \sim Po(\gamma \int_0^t \delta \exp(\delta s) ds) = Po(\gamma \cdot F(t)),$$

where $\gamma = \beta/\delta$, and $F(t) = 1 - \exp(-\delta t)$, the cumulative density function of exponential (δ). From this we deduce that the popu-

lation size $N(\infty) = Po(\gamma)$. Furthermore, the survivors' age distribution is an empirical distribution of a sample of exponentially distributed random variables, generated with the parameter δ.

To obtain an estimate of the true age, we must consider the error of K_s with respect to the true age of paralogs. If the true age is T, then we can calculate K_s (with error) as

$$K_s = T + (s|t)\ z,$$

where $s|t$ is the standard error for K_s at $T = t$, and z is a standard normal random variable. The error can be estimated from the empirical standard error given by the PAML software.

The mean of s is expected to correlate with the time t, since older K_s estimates have larger variances. The conditional distribution of s can be approximated by exponential $(2/t)$. The maximum likelihood estimate of the parameter δ from the data was obtained using a grid-based method, and a simulated sample under the null model was compared to the observed using a χ^2 test. A quantile–quantile plot (Q-Q plot) was used to visualize the difference between observed data and a simulated data set according to the null model. A strong deviation from the 45° line in the Q-Q plot suggests that the two distributions differ, and a bootstrapped Kolmogorov-Smirnov test (http://sekhon.polisci.berkeley.edu/matching/ks.boot.html) was applied to compare the observed and expected K_s distributions. The modeling and simulation scripts are available as Supplemental data.

Finite mixture model of genome duplications

In order to explore further how genome-wide duplication events influence the age distribution of paralogs and K_s distributions, we defined "background duplication" as gene duplication under the constant-rate birth–death process, and a "genome duplication" as an instant spike of gene birth overlaid on top of the background. We modeled changes in K_s distributions with increasing time since a duplication event, while assuming a constant rate of gene loss (death rate) and a constant background gene duplication rate (birth rate). Each simulation included a genome duplication (which led to new duplicates n) at time t. About 5% of duplicates were allowed to escape the death process.

In all instances when we rejected the constant rates hypothesis, we surmised that the observed K_s distributions actually reflect a compound distribution generated by variable birth and/or death rates from the time of duplication. For example, a genome duplication event would generate an immediate spike in the birth of paralogs. Mixture models treat the distribution of interest as a mixture of several component distributions in various proportions. The EMMIX software is suitable for mixed populations, where each component can be described by a Gaussian density (McLachlan et al. 1999) (see http://www.maths.uq.edu.au/~gjm/emmix/emmix.html for the Users' Guide). Following Schlueter et al. (2004), we modeled the log-transformed K_s distribution of paralogs. (The actual distribution is a mixture of log-transformed exponentials and normals.) Observations with $K_s < 0.005$ were excluded to avoid fitting a component to infinity (Schlueter et al. 2004). This truncation might also reduce the proportion of gene pairs attributed to background duplication. We modeled the mixed populations with one to four components and repeated the EM algorithm 100 times with random starting values, as well as 10 times with k-mean start values. One restriction imposed on the variance structure of K_s is that variance increases with the mean according to the empirical estimates. The observed data could therefore often be fitted to more than one component, with different means, variances, and mixture proportions. The mixture model with the best fit was identified using the Bayesian Information Criterion (Schwarz 1978). The mean and variance for each component (subpopulation of log K_s values) for the selected model were back-transformed to the original scale for plotting and interpretation.

Calibrating rate of synonymous substitution across lineages

When comparing K_s distributions among taxa, variation in the substitution rates among lineages must be taken into account. We used a phylogenetic approach to estimate lineage-specific synonymous substitution rates on branches leading to the magnoliids *L. tulipifera*, *P. americana*, and *S. henryi*. Orthologous genes from *A. thaliana*, rice, and the three magnoliid species were classified by InParanoid (Remm et al. 2001). Protein alignments of *Arabidopsis* and rice gene models (the TIGR *Arabidopsis thaliana* database, the TIGR rice database) were first constructed, then DNA alignments were forced to protein alignments by codon positions. A maximum likelihood tree was estimated using the HKY model in PHYML v.2.4.3 (Guindon and Gascuel 2003) for each putative ortholog set including at least 400 aligned nucleotide positions. A per-site estimate of K_s was then made for each magnoliid branch in gene phylogenies consistent with organismal relationships ([*Liriodendron*, *Persea*] *Saruma*) using codeml in the PAML package (Yang 1997). The ratio of K_s values on the *Persea* branch relative to the *Liriodendron* branch was then estimated for each gene.

Two supplemental tables and R-scripts for birth–death simulations are available as Supplemental material. Teri Solow and Lukas Muller provided the EST sequence assembly for eight species (*A. americanus, A. trichopoda, E. californica, L. tulipifera, N. advena, P. americana, S. henryi,* and *Welwitschia mirabilis*), now available through the Plant Genome Network (http://pgn.cornell.edu/).

Acknowledgments

We thank Jongmin Nam for providing code for K_s computation; Lena Scheaffer, Yi Hu, and Shelia Plock for technical support on cDNA library construction and sequencing; Lukas Mueller, Dan Ilut, Teri Solow, and Steve Tanksley for the PGN Database; and anonymous reviewers for critical comments on the manuscript. This work was supported by NSF Plant Genome award DBI-0115684.

References

Abi-Rached, L., Gilles, A., Shiina, T., Pontarotti, P., and Inoko, H. 2002. Evidence of en bloc duplication in vertebrate genomes. *Nat. Genet.* **31:** 100–105.

Adams, K.L., Cronn, R., Percifield, R., and Wendel, J.F. 2003. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci.* **100:** 4649–4654.

Albert, V.A., Soltis, D.E., Carlson, J.E., Farmerie, W.G., Wall, P.K., Ilut, D.C., Solow, T.M., Mueller, L.A., Landherr, L.L., Hu, Y., et al. 2005. Floral gene resources from basal angiosperms for comparative genomics research. *BMC Plant Biol.* **5:** 5.

Bell, C.D., Soltis, D.E., and Soltis, P.S. 2005. The age of the angiosperms: A molecular timescale without a clock. *Evolution Int. J. Org. Evolution* **59:** 1245–1258.

Bennett, M.D., Leitch, I.J., Price, H.J., and Johnston, J.S. 2003. Comparisons with *Caenorhabditis* (approximately 100 Mb) and *Drosophila* (approximately 175 Mb) using flow cytometry show genome size in *Arabidopsis* to be approximately 157 Mb and thus approximately 25% larger than the *Arabidopsis* Genome Initiative estimate of approximately 125 Mb. *Ann. Bot. (Lond.)* **91:** 547–557.

Bierne, N. and Eyre-Walker, A. 2003. The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: Implications for the correlation between the synonymous substitution rate and codon usage bias. *Genetics* **165:** 1587–1597.

Blanc, G. and Wolfe, K.H. 2004. Widespread paleopolyploidy in model

plant species inferred from age distribution of duplicate genes. *Plant Cell* **16:** 1667–1678.
Blanc, G., Barakat, A., Guyot, R., Cooke, R., and Delseny, M. 2000. Extensive duplication and reshuffling in the *Arabidopsis* genome. *Plant Cell* **12:** 1093–1101.
Blanc, G., Hokamp, K., and Wolfe, K.H. 2003. A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.* **13:** 137–144.
Bogart, J.P. 1979. Evolutionary implications of polyploidy in amphibians and reptiles. *Basic Life Sci.* **13:** 341–378.
Borsch, T., Hilu, K.W., Quandt, D., Wilde, V., Neinhuis, C., and Barthlott, W. 2003. Noncoding plastid trnT–trnF sequences reveal a well resolved phylogeny of basal angiosperms. *J. Evol. Biol.* **16:** 558–576.
Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433–438.
Brysting, A.K. and Borgen, L. 2000. Isozyme analysis of the *Cerastium alpinum* C-arcticum complex (Caryophyllaceae) supports a splitting of *C. arcticum* Lange. *Plant Syst. Evol.* **220:** 199–221.
Buzgo, M., Soltis, P.S., Kim, S., and Soltis, D.E. 2005. The making of a flower. *Biologist* **52:** 149–154.
Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D., and May, G. 2004. The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* **4:** 10.
Darlington, C.D. 1937. *Recent advances in cytology*. P. Blakiston's Son & Co., Philadelphia, PA.
Darwin, C.D. 1903. *More letters of Charles Darwin*. John Murray, London.
De Bodt, S., Maere, S., and Van de Peer, Y. 2005. Genome duplication and the origin of angiosperms. *Trends Ecol. Evol.* **20:** 591–597.
Dehal, P. and Boore, J.L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3:** e314.
deWet, J.M. 1979. Origins of polyploids. *Basic Life Sci.* **13:** 3–15.
Dong, Q., Schlueter, S.D., and Brendel, V. 2004. PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.* **32:** D354–D359.
Duvall, M.R., Learn Jr., G.H., Eguiarte, L.E., and Clegg, M.T. 1993. Phylogenetic analysis of rbcL sequences identifies *Acorus calamus* as the primal extant monocotyledon. *Proc. Natl. Acad. Sci.* **90:** 4641–4644.
Edgar, R.C. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5:** 113.
Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.
Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.
Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., and Postlethwait, J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151:** 1531–1545.
Friedman, R. and Hughes, A.L. 2001. Gene duplication and the structure of eukaryotic genomes. *Genome Res.* **11:** 373–381.
Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T. 1996. Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci.* **93:** 10274–10279.
Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725–736.
Grant, V. 1963. *The origin of adaptations*. Columbia University Press, New York.
———. 1981. *Plant speciation*. Columbia University Press, New York.
Grant, D., Cregan, P., and Shoemaker, R.C. 2000. Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. *Proc. Natl. Acad. Sci.* **97:** 4168–4173.
Gu, X., Wang, Y., and 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.
Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52:** 696–704.
Hilu, K.W. 1993. Polyploidy and the evolution of domesticated plants. *Am. J. Bot.* **80:** 2521–2528.
Hilu, K.W., Borsch, T., Mueller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M., Alice, L.A., Evans, R., et al. 2003. Angiosperm phylogeny based on matK sequence information. *Am. J. Bot.* **90:** 1758–1776.
Hughes, A.L. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. *J. Mol. Evol.* **48:** 565–576.
Hughes, A.L. and Friedman, R. 2003. 2R or not 2R: Testing hypotheses of genome duplication in early vertebrates. *J. Struct. Funct. Genomics* **3:** 85–93.
Iseli, C., Jongeneel, C.V., and Bucher, P. 1999. ESTScan: A program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 138–148.
Karev, G.P., Wolf, Y.I., Berezovskaya, F.S., and Koonin, E.V. 2004. Gene family evolution: An in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol. Biol.* **4:** 32.
Kellis, M., Birren, B.W., and Lander, E.S. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428:** 617–624.
Kim, S., Yoo, M.-J., Albert, V.A., Farris, J.S., Soltis, P.S., and Soltis, D.E. 2004. Phylogeny and diversification of B-function MADS-box genes in angiosperms: Evolutionary and functional implications of a 260-million-year-old duplication. *Am. J. Bot.* **91:** 2102–2118.
Kramer, E.M. and Irish, V.F. 1999. Evolution of genetic mechanisms controlling petal development. *Nature* **399:** 144–148.
Kramer, E.M., Dorit, R.L., and Irish, V.F. 1998. Molecular evolution of genes controlling petal and stamen development: Duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. *Genetics* **149:** 765–783.
Ku, H.M., Vision, T., Liu, J., and Tanksley, S.D. 2000. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci.* **97:** 9121–9126.
Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K., and dePamphilis, C.W. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. *Mol. Biol. Evol.* **22:** 1948–1963.
Li, W.H. and Grauer, D. 1991. *Fundamentals of molecular evolution*. Sinauer Associates, Sunderland, MA.
Liu, B. and Wendel, J.F. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol. Phylogenet. Evol.* **29:** 365–379.
Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.
Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M., and Van de Peer, Y. 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci.* **102:** 5454–5459.
Makalowski, W. 2001. Are we polyploids? A brief history of one hypothesis. *Genome Res.* **11:** 667–670.
Masterson, J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. *Science* **264:** 421–424.
McLachlan, G.J., Peel, D., Basford, K.E., and Adams, P. 1999. The EMMIX software for the fitting of mixtures of normal and t-components. *J. Stat. Softw.* **4:** 2.
McLysaght, A., Hokamp, K., and Wolfe, K.H. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* **31:** 200–204.
Müntzing, A. 1936. The evolutionary significance of autopolyploidy. *Hereditas* **21:** 263–378.
Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.
Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York.
Otto, S.P. and Whitton, J. 2000. Polyploid incidence and evolution. *Annu. Rev. Genet.* **34:** 401–437.
Paterson, A.H., Bowers, J.E., and Chapman, B.A. 2004a. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl. Acad. Sci.* **101:** 9903–9908.
Paterson, A.H., Bowers, J.E., Chapman, B.A., Peterson, D.G., Rong, J., and Wicker, T.M. 2004b. Comparative genome analysis of monocots and dicots, toward characterization of angiosperm diversity. *Curr. Opin. Biotechnol.* **15:** 120–125.
Remm, M., Storm, C.E., and Sonnhammer, E.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* **314:** 1041–1052.
Schlueter, J.A., Dixon, P., Granger, C., Grant, D., Clark, L., Doyle, J.J., and Shoemaker, R.C. 2004. Mining EST databases to resolve evolutionary events in major crop species. *Genome* **47:** 868–876.
Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Statist.* **6:** 461–464.
Shoemaker, R.C., Polzin, K., Labate, J., Specht, J., Brummer, E.C., Olson, T., Young, N., Concibido, V., Wilcox, J., Tamulonis, J.P., et al. 1996. Genome duplication in soybean (*Glycine* subgenus *soja*). *Genetics* **144:** 329–338.
Simillion, C., Vandepoele, K., Van Montagu, M.C., Zabeau, M., and Van de Peer, Y. 2002. The hidden duplication past of *Arabidopsis thaliana*.

Proc. Natl. Acad. Sci. **99:** 13627–13632.
Soltis, D.E. and Soltis, P.S. 1990. Isozyme evidence for ancient polyploidy in primitive angiosperms. *Syst. Bot.* **15:** 328–337.
———. 1999. Polyploidy: Recurrent formation and genome evolution. *Trends Ecol. Evol.* **14:** 348–352.
Soltis, P.S., Soltis, D.E., Zanis, M.J., and Kim, S. 2000. Basal lineages of angiosperms: Relationships and implications for floral evolution. *Int. J. Plant Sci.* **161:** S97–S107.
Soltis, D.E., Soltis, P.S., and Zanis, M.J. 2002. Phylogeny of seed plants based on evidence from eight genes. *Am. J. Bot.* **89:** 1670–1681.
Stebbins, G.L. 1950. *Variation and evolution in plants*. Columbia University Press, New York.
Stefanovic, S., Rice, D.W., and Palmer, J.D. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol. Biol.* **4:** 35.
Vision, T.J., Brown, D.G., and Tanksley, S.D. 2000. The origins of genomic duplications in *Arabidopsis*. *Science* **290:** 2114–2117.
Wang, H.C., Singer, G.A., and Hickey, D.A. 2004a. Mutational bias affects protein evolution in flowering plants. *Mol. Biol. Evol.* **21:** 90–96.
Wang, J.P., Lindsay, B.G., Leebens-Mack, J., Cui, L., Wall, K., Miller, W.C., and dePamphilis, C.W. 2004b. EST clustering error evaluation and correction. *Bioinformatics* **20:** 2973–2984.
Wang, W., Tanurdzic, M., Luo, M., Sisneros, N., Kim, H.R., Weng, J.K., Kudrna, D., Mueller, C., Arumuganathan, K., Carlson, J., et al. 2005. Construction of a bacterial artificial chromosome library from the spikemoss *Selaginella moellendorffii*: A new resource for plant comparative genomics. *BMC Plant Biol.* **5:** 10.
Wolfe, K.H. and Shields, D.C. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387:** 708–713.
Wolfe, K.H., Li, W.H., and Sharp, P.M. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci.* **84:** 9054–9058.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17:** 32–43.
Yu, J.J., Wang, W., Lin, S., Li, H., Li, J., Zhou, P., Ni, W., Dong, S., Hu, C., Zeng, J., et al. 2005. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3:** e38.
Zahn, L.M., Kong, H., Leebens-Mack, J.H., Kim, S., Soltis, P.S., Landherr, L.L., Soltis, D.E., dePamphilis, C.W., and Ma, H. 2005a. The evolution of the SEPALLATA subfamily of MADS-box genes: A preangiosperm origin with multiple duplications throughout angiosperm history. *Genetics* **169:** 2209–2223.
Zahn, L.M., Leebens-Mack, J., dePamphilis, C.W., Ma, H., and Theissen, G. 2005b. To B or Not to B a flower: The role of DEFICIENS and GLOBOSA orthologs in the evolution of the angiosperms. *J. Hered.* **96:** 225–240.
Zanis, M.J., Soltis, D.E., Soltis, P.S., Mathews, S., and Donoghue, M.J. 2002. The root of the angiosperms revisited. *Proc. Natl. Acad. Sci.* **99:** 6848–6853.
Zhang, J., Rosenberg, H.F., and Nei, M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci.* **95:** 3708–3713.

Received October 20, 2005; accepted in revised form March 27, 2006.