Syst. Biol. 0(0):1–14, 2014

DOI:10.1093/sysbio/syu055

Coalescent versus Concatenation Methods and the Placement of *Amborella* as Sister to Water Lilies

ZHENXIANG XI[1], LIANG LIU[2], JOSHUA S. REST[3], AND CHARLES C. DAVIS[1,*]

[1]*Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA;* [2]*Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA; and* [3]*Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794, USA;*

[*]*Correspondence to be sent to: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; E-mail: cdavis@oeb.harvard.edu.*

Received 7 January 2014; reviews returned 2 April 2014; accepted 24 July 2014
Associate Editor: Erika Edwards

Abstract.—The molecular era has fundamentally reshaped our knowledge of the evolution and diversification of angiosperms. One outstanding question is the phylogenetic placement of *Amborella trichopoda* Baill., commonly thought to represent the first lineage of extant angiosperms. Here, we leverage publicly available data and provide a broad coalescent-based species tree estimation of 45 seed plants. By incorporating 310 nuclear genes, our coalescent analyses strongly support a clade containing *Amborella* plus water lilies (i.e., Nymphaeales) that is sister to all other angiosperms across different nucleotide rate partitions. Our results also show that commonly applied concatenation methods produce strongly supported, but incongruent placements of *Amborella*: slow-evolving nucleotide sites corroborate results from coalescent analyses, whereas fast-evolving sites place *Amborella* alone as the first lineage of extant angiosperms. We further explored the performance of coalescent versus concatenation methods using nucleotide sequences simulated on (i) the two alternate placements of *Amborella* with branch lengths and substitution model parameters estimated from each of the 310 nuclear genes and (ii) three hypothetical species trees that are topologically identical except with respect to the degree of deep coalescence and branch lengths. Our results collectively suggest that the *Amborella* alone placement inferred using concatenation methods is likely misled by fast-evolving sites. This appears to be exacerbated by the combination of long branches in stem group angiosperms, *Amborella*, and Nymphaeales with the short internal branch separating *Amborella* and Nymphaeales. In contrast, coalescent methods appear to be more robust to elevated substitution rates. [*Amborella trichopoda*; coalescent methods; concatenation methods; elevated substitution rates; long-branch attraction; Nymphaeales.]

Angiosperms are the most diverse plant clade in modern terrestrial ecosystems. Although tremendous progress has been made clarifying their origins and diversification, one outstanding question is the early branching order of extant angiosperms, especially the phylogenetic placement of the New Caledonian endemic *Amborella trichopoda* Baill. Numerous studies concatenating multiple genes with dense taxon sampling have independently converged on *Amborella* as the lone sister to all other extant angiosperms (Parkinson et al. 1999; Qiu et al. 1999, 2000, 2005; Soltis et al. 1999, 2000, 2011; Zanis et al. 2002; Zhang et al. 2012; Drew et al. 2014). In addition, one study using duplicate gene rooting has similarly supported this hypothesis (Mathews and Donoghue 1999). However, a smaller number of studies using additional genes, especially slowly evolving genes, and analyses that more exhaustively handle rate heterogeneity have suggested that a clade containing *Amborella* plus water lilies (i.e., Nymphaeales) cannot be excluded as the first lineage of extant angiosperms (Barkman et al. 2000; Stefanović et al. 2004; Leebens-Mack et al. 2005; Soltis et al. 2007; Finet et al. 2010; Qiu et al. 2010; Wodniok et al. 2011). In particular, attempts to systematically remove fast-evolving sites that are more prone to saturation due to high rates of nucleotide substitution have led to increased support for the *Amborella* plus Nymphaeales hypothesis (Goremykin et al. 2009, 2013; Drew et al. 2014). A broader comparative phylogenomic assessment of this question is needed to better understand the placement of *Amborella*, and this is especially timely in light of the recent publication of its genome (Albert et al. 2013).

Advances in next-generation sequencing and computational phylogenomics represent tremendous opportunities for inferring species relationships using hundreds, or even thousands, of genes. Until now the reconstruction of broad angiosperm phylogenies from multiple genes has relied almost entirely on concatenation methods (Parkinson et al. 1999; Qiu et al. 1999, 2000, 2005; Soltis et al. 1999, 2000, 2011; Jansen et al. 2007; Moore et al. 2007, 2010, 2011; Wang et al. 2009; Lee et al. 2011; Zhang et al. 2012; Drew et al. 2014), in which phylogenies are inferred from a single combined gene matrix (Huelsenbeck et al. 1996). These analyses assume that all genes have the same, or very similar, evolutionary histories. Theoretical and simulation studies, however, have shown that concatenation methods can yield misleading results, especially if the true species tree is in an "anomaly zone" (Kubatko and Degnan 2007; Liu and Edwards 2009). This region of branch length space is characterized by a set of short internal branches in which the most frequently produced gene tree differs from the topology of the species tree (Degnan and Rosenberg 2006; Kubatko and Degnan 2007; Rosenberg and Tao 2008; Liu and Edwards 2009). Importantly, the boundaries of the anomaly zone can be expanded with uncertainty in gene tree estimation due to the random process of mutation (Huang and Knowles 2009). Recently developed coalescent-based methods permit gene trees to have different evolutionary histories

(Rannala and Yang 2003; Liu and Pearl 2007; Kubatko et al. 2009; Liu et al. 2009a, 2009b, 2010; Heled and Drummond 2010; Wu 2012), and both theoretical and empirical studies have demonstrated that coalescent methods better accommodate topological heterogeneity among gene trees (Liu et al. 2009b, 2010; Song et al. 2012; Zhong et al. 2013). Moreover, one recent study has hypothesized that coalescent methods might also reduce the potential deleterious effect of elevated substitution rates in phylogenomic analyses (Xi et al. 2013), but this has not been more thoroughly investigated.

Here, we leverage publicly available data from whole-genome sequencing projects and deeply sequenced transcriptomes to investigate the earliest diverging lineage of extant angiosperms. By incorporating hundreds of nuclear genes, we provide a direct comparison of phylogenetic relationships inferred among sites with different substitution rates using both coalescent and concatenation methods.

Materials and Methods

Data Acquisition and Sequence Translation

Gene sequences from both nuclear and plastid genomes were assembled using publicly available data. Our nuclear gene taxon sampling included 42 species representing all major angiosperm clades (35 families and 28 orders *sensu* Bremer et al. [2009]; Supplementary Table S1, available from http://dx.doi.org/10.5061/dryad.qb251). Three gymnosperms (*Picea glauca* [Moench] Voss, *Pinus taeda* L., and *Zamia vazquezii* D.W. Stev., Sabato & De Luca) and one lycophyte (*Selaginella moellendorffii* Hieron.) were included as outgroups. These three gymnosperms span the crown node of extant gymnosperms (Xi et al. 2013). Coding sequences were acquired for 25 species from whole-genome sequencing projects (Supplementary Table S1); for the remaining 21 species, assembled transcripts were obtained from PlantGDB (Duvick et al. 2008) and the Ancestral Angiosperm Genome Project (Jiao et al. 2011), and translated to amino acid sequences using prot4EST v2.2 (Wasmuth and Blaxter 2004).

To compare the evolutionary history between nuclear and plastid genomes, we obtained the annotated plastid genomes from GenBank for 37 angiosperm species (Supplementary Table S2), plus three gymnosperms (*Picea morrisonicola* Hayata, *Pinus koraiensis* Siebold & Zucc., and *Cycas taitungensis* Shen, Hill, Tsou, & Chen) and one lycophyte (*S. moellendorffii*) as outgroups. These 41 species represent the same taxonomic orders as those in our nuclear gene analyses.

Homology Assignment and Sequence Alignment

The establishment of sequence homology for both nuclear and plastid genes followed Dunn et al. (2008) and Hejnol et al. (2009). Briefly, sequence similarity was first assessed for all amino acid sequences using BLASTP v2.2.25 (Altschul et al. 1990) with $10^{-20}e$ -value threshold, and then grouped with MCL v09-308 using a Markov cluster algorithm (Enright et al. 2002). Each gene cluster was required to (i) include at least one sequence from *Selaginella* (for outgroup rooting), (ii) include sequences from at least four species, (iii) include at least 100 amino acids for each sequence following Liu and Xue (2005), (iv) have a mean of less than five homologous sequences per species, and (v) have a median of less than two sequences per species. Amino acid sequences from each gene cluster were aligned using MUSCLE v3.8.31 (Edgar 2004), and ambiguous sites were trimmed using trimAl v1.2rev59 (Capella-Gutiérrez et al. 2009) with the heuristic automated method. Sequences were removed from the alignment if they contained less than 70% of the total alignment length (Jiao et al. 2012). Nucleotide sequences were then aligned according to the corresponding amino acid alignments using PAL2NAL v14 (Suyama et al. 2006). For each gene cluster, the best-scoring maximum-likelihood (ML) tree was inferred from nucleotide alignments using RAxML v7.2.8 (Stamatakis 2006) with the GTRGAMMA substitution model, and rooted with *Selaginella*. All but one sequence were deleted in clades of sequences derived from the same species (i.e., monophyly masking) using Phyutility v2.2.6 (Smith and Dunn 2008).

Paralog Pruning and Species Tree Assessment

To reduce the potential negative effect of gene duplication and gene loss in inferring phylogenetic relationships from nuclear genes, especially for early diverging angiosperms, we further (i) excluded those gene clusters with paralogs associated with genome duplications in the common ancestor of extant seed plants and angiosperms identified by Jiao et al. (2011), (ii) included only those gene clusters containing one sequence from *Amborella* and one from Nymphaeales (i.e., *Nuphar advena* [Aiton] W.T. Aiton), and (iii) eliminated paralogs from more recent duplications (e.g., polyploidy associated with core eudicots [Jiao et al. 2012], legumes [Pfeil et al. 2005; Bertioli et al. 2009], monocots [Tang et al. 2010], and mustards [Bowers et al. 2003]) in each gene cluster using the paralog pruning described by Hejnol et al. (2009). Using this paralog pruning, we identified the maximally inclusive subtree in each gene tree, which contains no more than one sequence per species. Subtrees were then filtered to include only those with (i) 16 or more species and (ii) 60% of the species present in the original gene cluster from which they were derived. In this manner, we more effectively balanced comprehensive taxon with comprehensive character sampling.

Species relationships were first estimated from nuclear gene trees using two recently developed coalescent methods: Species Tree Estimation using Average Ranks of Coalescence (STAR) (Liu et al. 2009b) as implemented in Phybase v1.3 (Liu and Yu 2010) and Maximum Pseudo-likelihood for Estimating Species Trees (MP-EST) v1.4 (Liu et al. 2010). Since both methods are

based on summary statistics calculated across all gene trees, a small number of outlier genes that significantly deviate from the coalescent model have relatively little effect on the ability of these methods to accurately infer the species trees (Song et al. 2012). We compared results from coalescent analyses of nuclear genes with those from concatenation analyses. The concatenated nuclear and plastid matrices were generated from individual genes using Phyutility. For ML analyses, the ML trees were inferred from each concatenated nucleotide matrix using RAxML with two partitioning strategies: OnePart (a single partition with the GTRGAMMA model) and GenePart (partitioned *a priori* by gene with a GTRGAMMA model for each partition). Bootstrap support was estimated using a multilocus bootstrapping approach (Seo 2008) with 200 replicates. The Bayesian analyses were performed using PhyloBayes MPI v1.4e (Lartillot et al. 2013) under the CAT–GTR model (Lartillot and Philippe 2004), which accounts for across-site rate heterogeneity using an infinite mixture model. Two independent Markov chain Monte Carlo (MCMC) analyses were conducted for each concatenated nucleotide matrix. Each MCMC analysis was run for 5000 cycles with trees being sampled every cycle, and the consistency of likelihood values and estimated parameter values from two MCMC analyses was determined using Tracer v1.5. Bayesian posterior probabilities (PPs) were calculated by building a 50% majority rule consensus tree from two MCMC analyses after discarding the 20% burn-in samples.

Alternative topology tests were performed in a ML framework using the approximately unbiased (AU) test (Shimodaira 2002). In each case, the alternative placement of *Amborella* was enforced, and the constrained searches were conducted using RAxML with OnePart for the concatenated nucleotide matrix. This constrained ML tree was then tested against the unconstrained ML tree using scaleboot v0.3-3 (Shimodaira 2008).

Estimation of Evolutionary Rate and Nucleotide Substitution Saturation

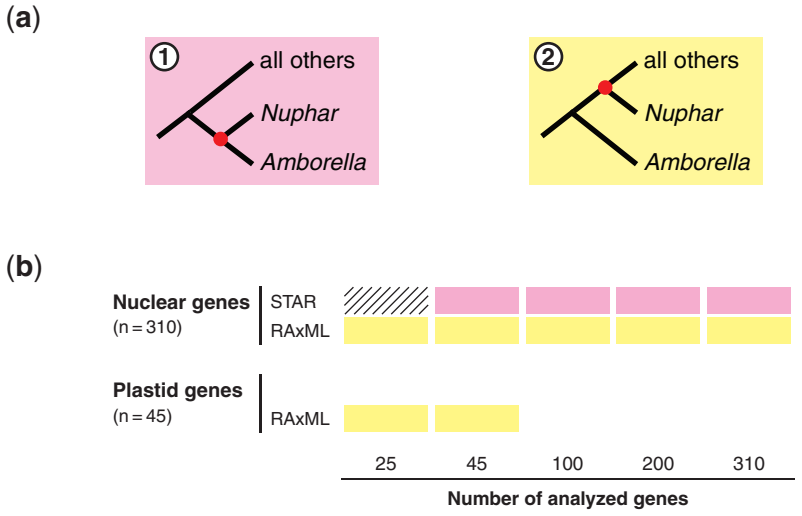
To evaluate the effect of elevated substitution rates for nuclear and plastid genes, we estimated the relative evolutionary rate for each of the nucleotide sites in our concatenated matrices using the observed variability (OV) (Goremykin et al. 2010) and Tree Independent Generation of Evolutionary Rates (TIGER) (Cummins and McInerney 2011) methods. The OV method calculates the total number of pair-wise mismatches at a given site, whereas the TIGER method uses similarity in the pattern of character-state distributions between sites as a proxy for site variability. Importantly, both OV and TIGER are tree-independent approaches. Thus, they are free from any systematic bias in estimating evolutionary rates attributable to an inaccurate phylogeny (Goremykin et al. 2010; Cummins and McInerney 2011).

We initially ignored parsimony uninformative sites and sorted all parsimony informative sites in our concatenated matrices based on their estimated evolutionary rates. We then divided these parsimony informative sites into two equal rate partitions—slow and fast. For the purpose of species tree estimation, we next redistributed these rate-classified sites back to their respective genes, effectively forming two subgenes from the original gene (i.e., the "slow" and "fast" subgenes). All parsimony uninformative sites from the same gene were included in both subgenes for proper model estimation. Species trees were then inferred from all "slow" subgenes and all "fast" subgenes separately. For coalescent analyses, individual gene trees were inferred using RAxML with the GTRGAMMA model, and rooted with *Selaginella*. These estimated gene trees were then used to construct the species trees with STAR and MP-EST. For concatenation analyses, the ML trees were inferred using RAxML with OnePart.

For each rate partition, nucleotide substitution saturation was measured using an entropy-based index of substitution saturation (I_{SS}) (Xia et al. 2003) as implemented in DAMBE (Xia and Xie 2001). I_{SS} was estimated for each rate partition from 200 replicates with gaps treated as unknown states. To reduce the effect of base compositional heterogeneity (Foster 2004), species relationships and bootstrap support were also estimated from concatenated nucleotide matrices using a nonhomogeneous, nonstationary model of DNA sequence evolution (Galtier and Gouy 1998; Boussau and Gouy 2006) as implemented in nhPhyML with default settings.

Simulation of Nucleotide Sequences to Evaluate the Two Alternative Placements of Amborella

To further evaluate the effect of elevated substitution rates on the placement of *Amborella*, we simulated nucleotide sequences based on the two alternative placements of this species (Fig. 1a). For each simulation, "*X*" percent of the 310 nuclear genes (where "*X*" ranges from 0 to 100 in increments of 10) were randomly assigned topology 1 (i.e., *Amborella* + *Nuphar* as the first lineage of angiosperms; Fig. 2), and the remaining genes were assigned topology 2 (i.e., *Amborella* alone as the first lineage of angiosperms; Supplementary Fig. S1). For each nuclear gene, the branch lengths of the assigned topology and parameters of the GTRGAMMA model were estimated from the original nucleotide sequences using RAxML with the "-f e" option. The resulting optimized gene tree and model parameters were then utilized to simulate nucleotide sequences using Seq-Gen v1.3.3 (Rambaut and Grassly 1997) with the GTR + Γ4 model. The concatenated nucleotide matrix was next generated from these 310 simulated genes using Phyutility. Sites were then sorted using the OV method and divided into slow and fast rate partitions as described above. Next, species trees were inferred for each rate partition using STAR, MP-EST,



(c)

Data	Sorting criteria	Rate partitions	No. of sites	Missing data	I_{SS}	Coalescence		Concatenation	
						STAR	MP-EST	RAxML	nhPhyML
Nuclear genes (46 species)	OV	*slow*	71,295	29.3%	0.728#	100	98	90	90
		fast	71,295	31.2%	0.918	58	93	99	99
	TIGER	*slow*	71,295	32.5%	0.764#	100	100	95	99
		fast	71,295	27.9%	0.883		96	100	100
Plastid genes (41 species)	OV	*slow*	10,199	4.1%	0.310§	n/a	n/a	100	99
		fast	10,199	4.1%	0.571#	n/a	n/a	99	100
	TIGER	*slow*	10,199	3.2%	0.308§	n/a	n/a	98	97
		fast	10,199	5.0%	0.572#	n/a	n/a	98	100

FIGURE 1. The placements of *Amborella trichopoda* inferred from our empirical data using different gene subsampling and nucleotide rate partitions. a) Two alternative placements of *Amborella*. For each placement, the red dot highlights the node of interest, and for which BP support is discussed in the main text. b) Support for the two alternative placements of *Amborella* inferred from coalescent (STAR) and concatenation (RAxML) analyses across subsampled gene categories. The 310 nuclear genes were subsampled for four different gene size categories (i.e., 25, 45, 100, and 200 genes; 10 replicates each), and the 45 plastid genes were subsampled for 25 genes (10 replicates). Cell with hatching indicates that support for the placement of *Amborella* from all replicates is below 80 BP; colored cells (pink = *Amborella* + *Nuphar*, yellow = *Amborella* alone) indicate relationships that received bootstrap support ≥80 BP from at least one replicate. c) Support for the two alternative placements of *Amborella* inferred from coalescent (STAR and MP-EST) and concatenation (RAxML and nhPhyML) analyses across different nucleotide rate partitions. Sites in each data set were sorted by evolutionary rates determined using the OV or TIGER method, and divided into two equal partitions (i.e., slow and fast). The index of substitution saturation (I_{SS}) was estimated from 32 terminals (§= I_{SS} is significantly smaller than the critical I_{SS} value [$I_{SS.C}$] when the true topology is pectinate or symmetrical; # = I_{SS} is significantly smaller than $I_{SS.C}$ when the true topology is symmetrical, but not significantly smaller than $I_{SS.C}$ when the true topology is pectinate; see Supplementary Table S7 for full results). Cell with hatching indicates that support for the placement of *Amborella* is below 50 BP.

and RAxML as described above. Each simulation was repeated 100 times.

Simulation of Nucleotide Sequences under the Coalescent Model

To more generally examine the effect of elevated substitution rates and discordant gene tree topologies independent of the *Amborella* data, we simulated gene trees using three hypothetical six-taxon species trees under a multispecies coalescent model (Rannala and Yang 2003). These three species trees (Fig. 3a) are topologically identical except with respect to the degree of deep coalescence and branch lengths. In each of the species trees 6T-1, 6T-2, and 6T-3, species *A*–*E* are designated as ingroups, and the sixth species *F* is designated as the outgroup. The branch lengths of the four internal branches in three species trees, a1 = a2 = a3 = a4 = 0.001, were held constant (branch lengths are in mutation units, i.e., the number of substitutions per site). The branch lengths of the external branches leading to species *A* and *F* (i.e., b1 = 0.001 and b6 = 0.004, respectively) were also held constant in all species trees. Thus, these three species trees differ only in the branch lengths of the four external branches leading to species *B*–*E* (i.e., b2, b3, b4, and b5, respectively), which we varied to simulate elevated nucleotide substitution rates. For the species tree 6T-1, branch lengths of the four external branches are: b2 = b3 = b4 = 0.001 and b5 = 0.003;

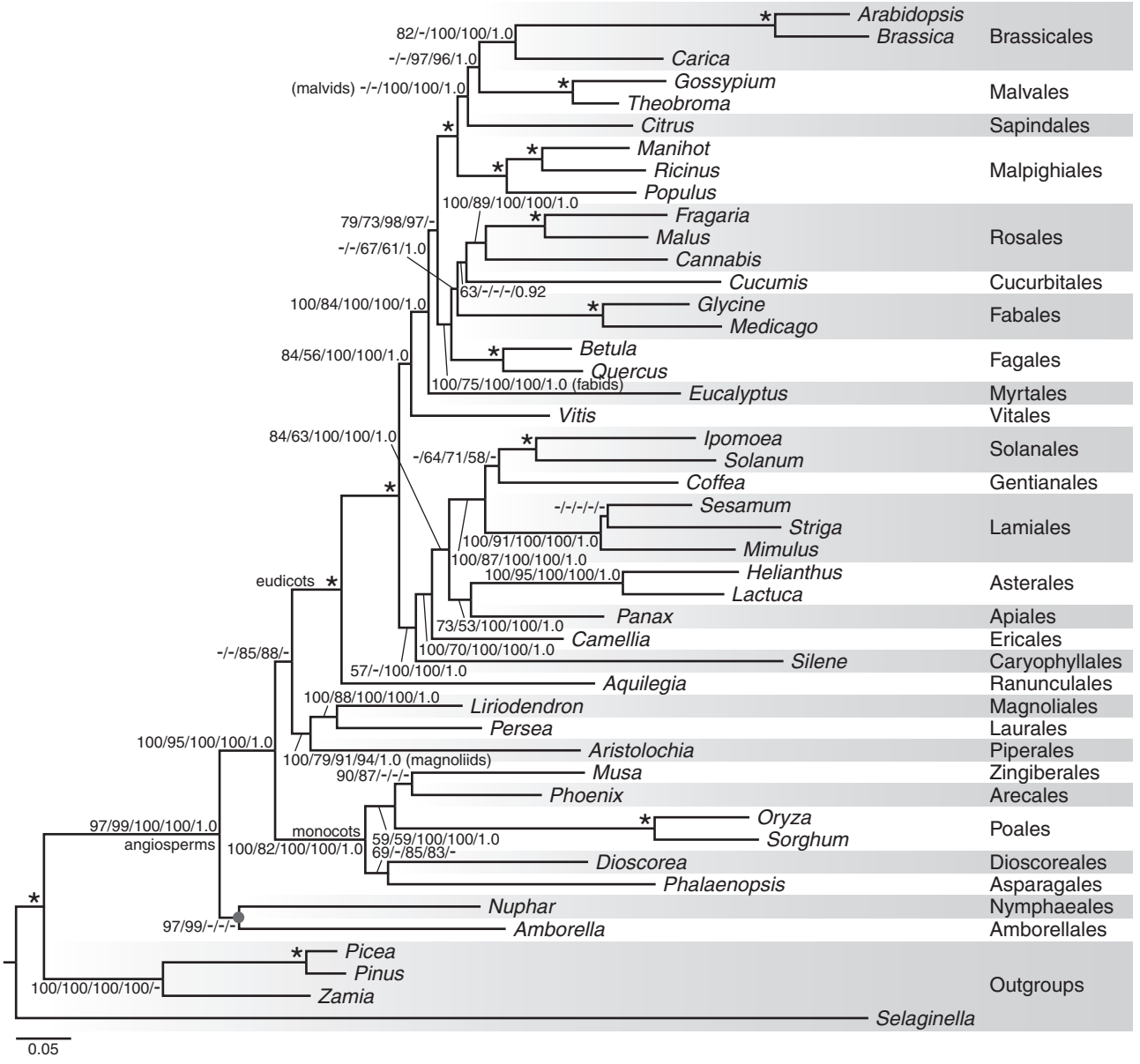


FIGURE 2. Species tree inferred from our 310 nuclear genes using the coalescent method (STAR). BPs and PP from STAR/MP-EST/RAxML with OnePart/RAxML with GenePart/PhyloBayes are indicated above each branch; an asterisk indicates that the clade is supported by 100 BPs and 1.0 PP from STAR, MP-EST, RAxML, and PhyloBayes. Branch lengths shown here were estimated for the concatenated nuclear matrix using RAxML with OnePart.

and for the species trees 6T-2 and 6T-3, branch lengths are: b2 = b3 = b4 = 0.101 and b5 = 0.103. For this simulation, we assumed that each gene lineage simulated from a branch in the species tree was subject to the same substitution rate specified for that branch. Thus, all gene trees simulated on species trees 6T-2 and 6T-3 possess longer external branches leading to species *B–E* compared with gene trees simulated on the species tree 6T-1.

In addition, each species tree has the same population size for all internal branches. Here, the population size parameter is defined as $\theta = 4\mu N_e$, where N_e is the effective population size and μ is the average mutation rate per site per generation (Liu and Yu 2010). We applied two different values of θ to simulate varying degrees of deep coalescence (i.e., $\theta = 0.0001$ for the species tree 6T-2 and $\theta = 0.01$ for the species trees 6T-1 and 6T-3). According to coalescent theory, the amount of deep coalescence is positively correlated with the value of θ, and a large value of θ produces gene trees with highly variable topologies despite a common species tree. Since these three species trees have the same branch length for internal branches (i.e., a1 = a2 = a3 = a4 = 0.001), the amount of deep coalescence depends only on the value of θ. Therefore, the species tree 6T-1 produced gene trees with highly discordant topologies (i.e., a

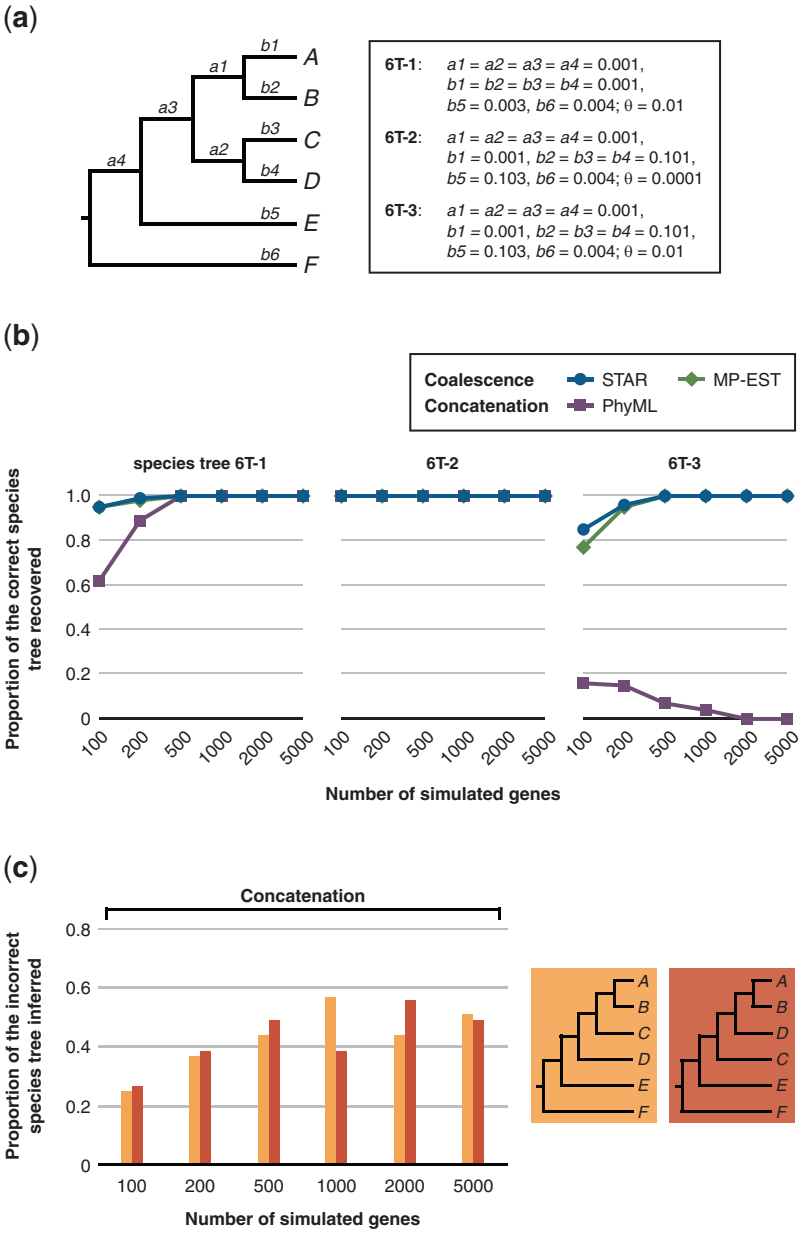


FIGURE 3. Performance of coalescent versus concatenation methods on nucleotide sequences simulated under the multispecies coalescent model. a) The topology and parameters of the three species trees 6T-1, 6T-2, and 6T-3 used to simulate nucleotide sequences. The branch lengths are in mutation units (i.e., the number of substitutions per site). The population size parameter is defined as $\theta = 4\mu N_e$, where N_e is the effective population size and μ is the average mutation rate per site per generation. b) Proportions of the correct species tree recovered by coalescent (STAR and MP-EST) and concatenation (PhyML) methods for each of the three species trees 6T-1, 6T-2, and 6T-3. c) Proportions of the two incorrect species trees inferred by the concatenation method from nucleotide sequences simulated on the species tree 6T-3.

high degree of deep coalescence), the species tree 6T-2 produced congruent gene trees (i.e., a low degree of deep coalescence) with long external branches, and the species tree 6T-3 produced gene trees with highly discordant topologies and long external branches.

We next simulated 100, 200, 500, 1000, 2000, and 5000 gene trees on each of the three species trees using Phybase, with one allele sampled from each species. Each gene tree was then utilized to simulate nucleotide sequences of 1000 bp using Seq-Gen with the JC69 model (Jukes and Cantor 1969). For coalescent analyses, since RAxML only allows the GTR model for nucleotide sequences, individual gene trees were inferred using PhyML v3.1 (Guindon et al. 2010) with the JC69 model, and rooted with species *F*. These estimated gene trees were then used to construct the species trees with STAR and MP-EST. For concatenation analyses, the ML trees were inferred for the concatenated nucleotide matrices using PhyML with the JC69 model. Each simulation was repeated 100 times.

Results and Discussion

Taxon and Gene Sampling of Nuclear and Plastid Genes

For nuclear genes, the approximately 1.4 million protein-coding sequences from the 46 species (Supplementary Table S1) were grouped into 19,101 gene clusters, 799 of which passed our initial criteria for selecting low-copy nuclear genes as described in the "Materials and Methods" section. Following this initial filter, the average numbers of sequences and species for each gene cluster were 32 and 30, respectively (Supplementary Fig. S2). Of these 799 gene clusters, 310 were retained for further phylogenetic analyses after paralog pruning, and the average number of species and nucleotide sites for each gene cluster were 33 and 773, respectively (Supplementary Table S3). The final concatenated nuclear matrix included 239,763 nucleotide sites (142,590 parsimony informative sites), 27.9% missing genes (Supplementary Table S4), and 29.9% missing data (including gaps).

For plastid genes, the 2172 protein-coding sequences from the 41 species (Supplementary Table S2) were grouped into 58 gene clusters, of which 45 remained following the filtering criteria described above. The average number of species and nucleotide sites for these 45 gene clusters were 40 and 1191, respectively (Supplementary Table S5). The final concatenated plastid matrix included 53,580 nucleotide sites (20,398 parsimony informative sites), 3.1% missing genes (Supplementary Table S6), and 4.9% missing data.

Inferring Species Relationships Using Coalescent versus Concatenation Methods

Our species trees inferred from nuclear and plastid genes largely agree with each other (Figs. 2 and 4). However, we identify four main conflicting relationships between the nuclear and plastid genomes. Our analyses of nuclear genes (Fig. 2) show that (i) monocots are sister to eudicots + magnoliids, (ii) Lamiales are sister to Gentianales + Solanales, (iii) Myrtales are sister to fabids + malvids, and (iv) Malpighiales are sister to malvids. In contrast, analyses of plastid genes (Fig. 4) show that (i) the magnoliids are sister to eudicots + monocots, (ii) Solanales are sister to Gentianales + Lamiales, (iii) Myrtales are sister to malvids, and (iv) Malpighiales are sister to the rest of fabids. These conflicting placements between the nuclear and plastid phylogenies are consistent with previous studies (e.g., Finet et al. 2010; Lee et al. 2011; Shulaev et al. 2011; Zhang et al. 2012), although ours is the first to include a balanced set of species and genes from both genomes. These results suggest that plastid and nuclear genomes have different evolutionary histories in several angiosperm clades.

The lone instance of strong discordance (≥80 bootstrap percentage [BP]) between the coalescent and concatenation analyses of nuclear genes is in the placement of *Amborella*. The coalescent analyses using STAR and MP-EST support a clade containing *Amborella* + *Nuphar* as the first angiosperm lineage with 97 and 99 BP, respectively (Fig. 2; see also red dots in Fig. 1a for nodes under consideration). In contrast, the concatenation analyses using RAxML and PhyloBayes place *Amborella* alone as the first lineage of angiosperms with 100/100 (OnePart/GenePart) BP and 1.0 PP, respectively. Similarly for plastid genes, the concatenation analyses using RAxML and PhyloBayes support *Amborella* alone as the first lineage with 83/82 BP and 1.0 PP, respectively (Fig. 4). Moreover, although the monophyly of *Amborella* + *Nuphar* cannot be rejected for our concatenated plastid matrix, it is rejected ($P < 0.001$) for the concatenated nuclear matrix using the AU test.

To further investigate if the placement of *Amborella* is sensitive to the number of sampled genes, we randomly subsampled our 310 nuclear genes in four different gene size categories (i.e., 25, 45, 100, and 200 genes; 10 replicates each). We similarly subsampled the 45 plastid genes (i.e., 25 genes with 10 replicates). Even as the sample size declines, the coalescent analyses (STAR) of the nuclear genes strongly support (≥80 BP) *Amborella* + *Nuphar* as the earliest diverging lineage of angiosperms. Support for this relationship only dropped below 80 BP when the number of subsampled nuclear genes was 25 (Fig. 1b). In contrast, the concatenation analyses (RAxML) strongly support (≥80 BP) *Amborella* alone as the first lineage in all gene sizes (Fig. 1b). Thus, the discordant placements of *Amborella* inferred from coalescent and concatenation analyses are robust to the number of genes sampled.

These analyses replicate the findings of many other genome-scale concatenation analyses that place *Amborella* alone as sister to all other extant angiosperms (e.g., Jansen et al. 2007; Moore et al. 2007, 2010; Lee et al. 2011), but ours is the first to show that coalescent analyses consistently and strongly support *Amborella* plus Nymphaeales together as the earliest diverging angiosperms.

Accommodating Elevated Rates of Substitution in Coalescent and Concatenation Analyses

It has long been appreciated that elevated rates of molecular evolution can lead to multiple substitutions at the same site (Olsen 1987; Salemi and Vandamme 2003; Goremykin et al. 2010). If the substitution model fails to effectively correct for high levels of saturation in fast-evolving sites, it could lead to the well-known phenomenon of long-branch attraction (LBA) (Felsenstein 1978). This can be especially prominent in resolving deeper relationships (Brinkmann and Philippe 1999; Hirt et al. 1999; Philippe et al. 2000; Gribaldo and Philippe 2002; Burleigh and Mathews 2004; Pisani 2004; Brinkmann et al. 2005; Goremykin et al. 2009, 2010; Philippe and Roure 2011; Zhong et al. 2011; Xi et al. 2013), and is likely to be relevant for inferring early angiosperm phylogeny given their ancient origin and well-documented rapid initial diversification (Wikström

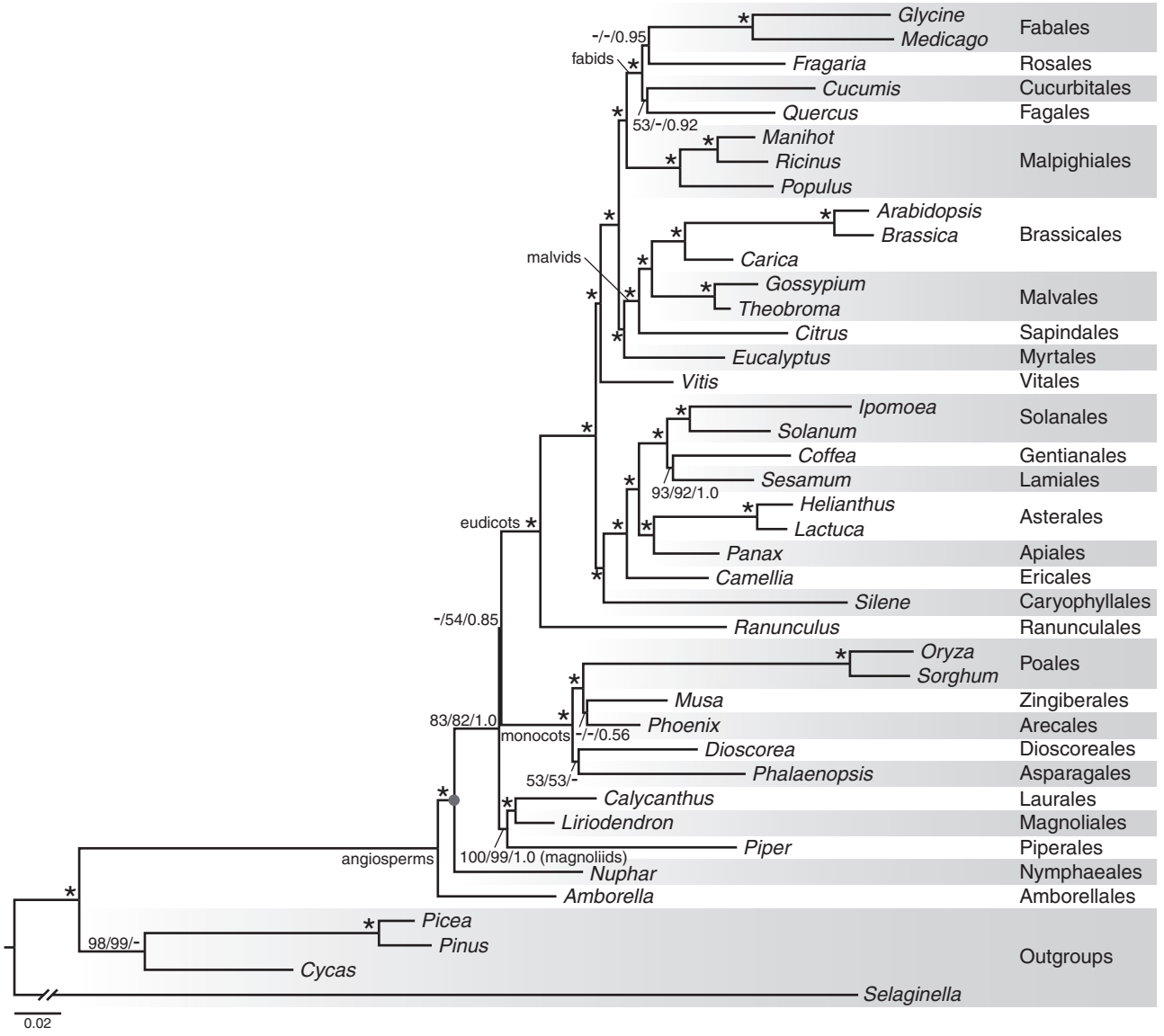


FIGURE 4. Species tree inferred from our 45 plastid genes using the concatenation method (RAxML). BPs and PP from RAxML with OnePart/RAxML with GenePart/PhyloBayes are indicated above each branch; an asterisk indicates that the clade is supported by 100 BPs and 1.0 PP from RAxML and PhyloBayes. Branch lengths shown here were estimated for the concatenated plastid matrix using RAxML with OnePart.

et al. 2001; Moore et al. 2007; Magallón and Castillo 2009; Bell et al. 2010; Smith et al. 2010). Some recent analyses using whole plastid genome data converge on the placement of *Amborella* as sister to Nymphaeales after identifying and removing fast-evolving sites in phylogenomic analyses (Goremykin et al. 2009, 2013; Drew et al. 2014). However, the effect of elevated substitution rates on angiosperm phylogeny has not been investigated broadly in the nuclear genome or with coalescent methods.

Here, we estimated the relative evolutionary rate for each of the sites in our concatenated nuclear and plastid matrices using the OV and TIGER methods (Supplementary Figs. S3 and S4), and examined the placement of *Amborella* for both slow and fast rate partitions. We find that the coalescent methods (STAR and MP-EST) support *Amborella* + *Nuphar* as the first lineage of extant angiosperms for both the slow (Supplementary Fig. S5) and fast (Supplementary Fig. S6) nuclear gene partitions. Support for this relationship drops below 80 BP only for the fast nuclear partition using STAR (Fig. 1c). In contrast, the concatenation method (RAxML) produces well supported but incongruent placements of *Amborella* across the two rate partitions for both the nuclear and plastid genes (Fig. 1c). Here, the slow nuclear gene (Supplementary Fig. S7) and slow plastid gene (Supplementary Fig. S8) partitions corroborate results from the coalescent analyses and strongly place (≥90 BP) *Amborella* + *Nuphar* as the first lineage of angiosperms. However, the fast partitions strongly support (≥98 BP) *Amborella* alone as the first lineage of angiosperms in all

Gene trees (n = 310)	OV - *slow*			OV - *fast*		
① : ②	STAR	MP-EST	RAxML	STAR	MP-EST	RAxML
0%:100%	1.0	1.0	1.0	1.0	0.89	1.0
10%:90%	1.0	0.99	1.0	1.0	0.98	1.0
20%:80%	1.0	1.0	1.0	1.0	0.95	1.0
30%:70%	1.0	0.97	1.0	0.99	0.90	1.0
40%:60%	0.97	0.97	1.0	0.93	0.84	1.0
50%:50%	0.66	0.60	0.98	0.59	0.51	1.0
60%:40%	1.0	0.95	0.90	0.80	0.87	1.0
70%:30%	1.0	1.0	1.0	1.0	0.99	1.0
80%:20%	1.0	1.0	1.0	0.99	1.0	0.91
90%:10%	1.0	1.0	1.0	1.0	1.0	0.67
100%:0%	1.0	1.0	1.0	1.0	1.0	0.99

FIGURE 5. Proportions of the two alternative placements of *Amborella trichopoda* recovered from simulated nuclear genes using coalescent (STAR and MP-EST) and concatenation (RAxML) methods. Nucleotide sequences were simulated on 310 nuclear gene trees representing varying percentages of the two alternative placements of *Amborella* (Fig. 1a) as indicated in the "Gene trees" column (pink = *Amborella* + *Nuphar*, yellow = *Amborella* alone). Sites in each data set were sorted by evolutionary rates determined using the OV method, and divided into two equal partitions (i.e., slow and fast).

nuclear gene (Supplementary Fig. S9) and plastid gene (Supplementary Fig. S10) analyses. Additionally, when the placement of *Amborella* + *Nuphar* is inferred using the concatenation method, the alternative placement of *Amborella* alone is rejected ($P < 0.05$, AU test). Similarly, in all cases when *Amborella* alone is supported, the alternative placement of *Amborella* + *Nuphar* is rejected ($P < 0.05$, AU test).

To determine whether nucleotide substitution saturation might influence the incongruent placements of *Amborella* in our concatenation analyses, we characterized sites within each rate partition using the index of substitution saturation (I_{SS}) (Xia et al. 2003). As I_{SS} approaches 1, or if I_{SS} is not smaller than the critical I_{SS} value ($I_{SS.C}$), then sequences are determined to exhibit substantial saturation (Xia et al. 2003). Our analyses demonstrate that for plastid genes (Fig. 1c and Supplementary Table S7), slow partitions exhibit no evidence of saturation (I_{SS} is significantly smaller than $I_{SS.C}$; $P < 0.001$, two-tailed *t*-test), whereas fast partitions show evidence of saturation (I_{SS} is not smaller than $I_{SS.C}$ when the true topology is pectinate with 32 terminals). In contrast, our analyses indicate that both rate partitions for nuclear genes show evidence of saturation (i.e., when the true topology is pectinate with 32 terminals; Fig. 1c and Supplementary Table S7), but slow partitions exhibit lower overall levels of saturation. To further minimize the influence of saturation, we selected the most conserved 5000 parsimony informative sites from our concatenated nuclear matrix. With this reduced data set, we no longer observe evidence of saturation ($P < 0.001$, two-tailed *t*-test; Supplementary Table S8), and the placement of *Amborella* + *Nuphar* is still supported with 93 BP using the concatenation method (RAxML). Thus, these results suggest that the incongruence we observe in the placement of *Amborella* across rate partitions using the concatenation method appears to be due to differences in the degree of nucleotide substitution saturation.

We next performed a simulation study to examine the effect of elevated substitution rates on the placement of *Amborella* in coalescent versus concatenation analyses. We used the branch lengths and nucleotide substitution parameters estimated from each of our 310 nuclear genes to simulate nucleotide sequences based on gene trees representing varying percentages of the two alternative placements of *Amborella* (Fig. 1a). Our results show that despite the discordant placements of *Amborella* in these simulated gene trees, the proportion of the correct placement of *Amborella* recovered by coalescent methods (STAR and MP-EST) is high, ranging from 0.80 to 1.0 for both slow and fast rate partitions (Fig. 5). For concatenation analyses (RAxML), when there is a single placement of *Amborella* in the simulated gene trees (i.e., "*X*" equals 0 or 100; see also the "Materials and Methods" section for details), despite rate heterogeneity across genes, the proportions of the correct placement of *Amborella* recovered by the concatenation method are very high (≥0.99) for both rate partitions (Fig. 5). In contrast, when 60–80% of genes are simulated with the *Amborella* + *Nuphar* topology enforced, the concatenation analyses produce incongruent placements of *Amborella* across the two rate partitions (Fig. 5). Here, the slow partitions again corroborate results from the coalescent analyses: the proportion of the correct placement of *Amborella* + *Nuphar* recovered by the concatenation method is high and ranges from 0.90 to 1.0. For fast partitions, however, the concatenation method infers the incorrect placement of *Amborella* alone at a very high rate (0.91–1.0). This observation that the concatenation analyses of fast partitions support the placement of *Amborella* alone, despite the fact that up to 80% of the genes are simulated with the alternative *Amborella* + *Nuphar*

topology, indicates that the concatenation analyses of fast partitions are biased toward the placement of *Amborella* alone even when it is incorrect. Therefore, this simulation indicates that analyzing data using coalescent methods, or only the slow partitions using concatenation methods, is more likely to recover the correct placement of *Amborella*. In addition, despite the fact that concatenation analyses of fast partitions recover the correct placement of *Amborella* + *Nuphar* at a very low rate of 0.09 when 80% of the genes are simulated with the *Amborella* + *Nuphar* topology (Fig. 5), on average 34.3% of the inferred gene trees still recover the correct placement of *Amborella* + *Nuphar* in fast partitions. This suggests that the negative effect of fast-evolving sites in ML analyses is more severe for concatenated gene sequences than for individual gene sequences.

We conducted a second simulation to examine the performance of coalescent versus concatenation methods under a multispecies coalescent model (Rannala and Yang 2003). These analyses were independent of the empirical data analyzed above, and were devised to investigate the influence of elevated substitution rates in particular lineages in combination with a high degree of deep coalescence. This is likely to be especially relevant to the placement of *Amborella* owing to the combination of long branches in stem group angiosperms, *Amborella*, and *Nuphar* with the short internal branch separating *Amborella* and *Nuphar* (Figs. 2 and 4). Our results of this simulation demonstrate that when nucleotide sequences were simulated on (i) gene trees with a high degree of deep coalescence but no long branches (i.e., for the species tree 6T-1 [Fig. 3a], on average 5.8% of the simulated gene trees matched the species tree topology) or (ii) gene trees with a low degree of deep coalescence but long external branches (i.e., for the species tree 6T-2 [Fig. 3a], all simulated gene trees matched the species tree topology), both coalescent (STAR and MP-EST) and concatenation (PhyML) methods accurately estimate the species tree as the number of genes increases (Fig. 3b). The proportion of the correct species tree recovered by both methods increases to 1.0 as the number of genes increases to 500 (Fig. 3b), indicating that both methods are not adversely affected when *either* discordant gene tree topologies owing to a high degree of deep coalescence *or* long external branches due to elevated substitution rates are present. In contrast, when nucleotide sequences were simulated on gene trees with *both* a high degree of deep coalescence *and* long external branches (i.e., for the species tree 6T-3 [Fig. 3a], on average 5.8% of the simulated gene trees matched the species tree topology), the coalescent methods still recover the correct species tree with a proportion of 1.0 as the number of genes increases to 500 (Fig. 3b). In contrast, the proportion of the correct species tree recovered by the concatenation method under these circumstances decreases to 0 as the number of genes increases to 2000 (Fig. 3b). Here, although nucleotide sequences simulated on the species tree 6T-3 show no evidence of saturation (the average I_{SS} equals 0.650, and the average $I_{SS.C}$ equals 0.794 when assuming a pectinate topology or 0.841 when assuming a symmetrical topology), the concatenation method consistently estimates two incorrect topologies: the long external branch leading to either species *C* or *D* is incorrectly attracted to the long external branch leading to species *E* (Fig. 3c). These simulation results strongly suggest that the combination of long external branches and short internal branches, especially when the degree of deep coalescence is high, may lead to the failure of concatenation methods. In contrast, coalescent methods appear to be more robust under these circumstances. Since the most probable gene tree matches the species tree 6T-3, our analyses further indicate that with elevated substitution rates, concatenation methods may consistently produce incorrect estimates even when the true species tree is not in the anomaly zone. Importantly, the pattern represented in the species tree 6T-3, in which an initial burst in diversification (i.e., short internal branches and a high degree of deep coalescence) is followed by long descendant branches of extant lineages, possibly characterizes numerous ancient rapid radiations across the Tree of Life (Whitfield and Lockhart 2007). Further study of this phenomenon will help to better understand the performance of coalescent versus concatenation methods under these circumstances.

In addition to the considerations raised above, recent studies have shown that base compositional heterogeneity can compromise phylogenetic analyses because commonly used substitution models assume equal nucleotide composition among taxa (Conant and Lewis 2001; Foster 2004; Jermiin et al. 2004; Sheffield et al. 2009; Nesnidal et al. 2010; Betancur et al. 2013). Here, we observed that the GC content of concatenated sequences ranged from 41.9% (*Aquilegia coerulea* James) to 53.7% (*Selaginella*) for nuclear genes (Supplementary Table S1) and from 37.2% (*Glycine max* [L.] Merr.) to 50.8% (*Selaginella*) for plastid genes (Supplementary Table S2). Therefore, as a further test, we analyzed our concatenated nucleotide matrices using a nonhomogeneous, nonstationary model of DNA sequence evolution (Galtier and Gouy 1998; Boussau and Gouy 2006) as implemented in nhPhyML. Our results here demonstrate that the slow partitions still place (≥90 BP) *Amborella* + *Nuphar* as the first lineage, whereas fast partitions support *Amborella* alone with ≥99 BP in all our nuclear and plastid analyses (Fig. 1c). Because this accommodation of base compositional heterogeneity does not change the incongruent placements of *Amborella* in concatenation analyses, we conclude that our results are not obviously influenced by variation in nucleotide base composition.

Finally, to confirm that the placement of *Amborella* as sister to Nymphaeales is not biased by insufficient taxon sampling that has been identified in earlier large-scale phylogenomic analyses (Soltis and Soltis 2004; Stefanović et al. 2004), we re-analyzed the recent 640-species 17-gene data set from Soltis et al. (2011) using the concatenation method (RAxML). This data set represents the broadest taxon and gene sampling to date for seed plants, and encompasses

330 families and 58 orders. It includes 17 genes representing all three plant genomic compartments (i.e., mitochondrion, nucleus, and plastid). The concatenation analyses mirror our phylogenomic results above. When analyzing only the slow partitions (7641 nucleotide sites; Supplementary Fig. S11), the clade containing *Amborella* plus Nymphaeales (*Brasenia*, *Cabomba*, *Nuphar*, *Nymphaea*, and *Trithuria*) is strongly supported as the first angiosperm lineage (91 BP and 94 BP for the slow OV and TIGER partitions, respectively). In contrast, when the fast partitions (7641 nucleotide sites) are analyzed, *Amborella* alone is inferred as the sister to all remaining angiosperms (78 BP and 80 BP for the fast OV and TIGER partitions, respectively).

Together with empirical and simulation results from above, our study indicates that the placement of *Amborella* alone inferred from concatenation analyses is likely misled by elevated nucleotide substitution rates. Moreover, given the combination of long branches in stem group angiosperms, *Amborella*, and Nymphaeales with the short internal branch separating *Amborella* and Nymphaeales (Figs. 2 and 4), this could be attributed to an LBA artifact involving fast-evolving sites. In contrast, coalescent methods appear to be more robust under these circumstances.

How does the placement of *Amborella* affect our understanding of early angiosperm evolution? One example is the egg apparatus. The female gametophyte of most extant angiosperms contains a three-celled egg apparatus at maturity (i.e., two synergids and an egg cell). One exception is *Amborella*, which possesses a unique four-celled egg apparatus (i.e., three synergids and an egg cell) (Friedman 2006). In an earlier phylogenetic reconstruction of the female gametophyte, when *Amborella* is placed as the lone sister to all other extant angiosperms it is equally parsimonious to hypothesize either the three- or four-celled egg apparatus as plesiomorphic in angiosperms (Friedman and Ryerson 2009). In contrast, our placement of *Amborella* as sister to Nymphaeales demonstrates that the common ancestor of angiosperms likely had a three-celled egg apparatus, and that the four-celled egg apparatus evolved independently in *Amborella*. Further ancestral state reconstructions are necessary to thoroughly understand additional aspects of early evolutionary history of angiosperms (cf. Barkman et al. 2000; Soltis et al. 2008; Doyle 2012; Doyle and Endress 2014).

The incongruence in concatenation analyses across sites with different evolutionary rates, which produce well supported, but conflicting placements of key taxa has also recently been reported in broader phylogenomic analyses of seed plants (Xi et al. 2013) and placental mammals (Song et al. 2012). In the case of seed plants, coalescent analyses consistently placed *Ginkgo* as sister to cycads; in the case of placental mammals, coalescent analyses demonstrated consistent and strong results for eutherian relationships, which were congruent with geographic data. Our results lend further empirical support for analyzing genome-scale data to resolve deep phylogenetic relationships using coalescent methods, and provide the most convincing evidence to date that *Amborella* plus Nymphaeales together represent the earliest diverging lineage of extant angiosperms. These results demonstrate that in the phylogenomic era, we not only need additional data to resolve difficult phylogenetic problems, but also sophisticated methods that reduce systematic errors in large-scale phylogenetic analyses (Philippe et al. 2011; Philippe and Roure 2011).

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad Digital Repository at http:// dx.doi.org/10.5061/dryad.qb251.

Funding

This work was supported by the United States National Science Foundation [DMS-1222745 to L.L. and DEB-1120243 to C.C.D.].

Acknowledgements

The authors thank Michael Donoghue, Dannie Durand, Peter Endress, and members of the Davis, Durand, and Rest laboratories for advice and discussion. They also thank Casey Dunn, Mike Ethier, and Alexandros Stamatakis for technical support. Finally, they thank the editors and two anonymous reviewers for their valuable comments and suggestions to improve the quality of the article.

References

Albert V.A., Barbazuk W.B., dePamphilis C.W., Der J.P., Leebens-Mack J., Ma H., Palmer J.D., Rounsley S., Sankoff D., Schuster S.C., Soltis D.E., Soltis P.S., Wessler S.R., Wing R.A., Ammiraju J.S.S., Chamala S., Chanderbali A.S., Determann R., Ralph P., Talag J., Tomsho L., Walts B., Wanke S., Chang T.H., Lan T.Y., Arikit S., Axtell M.J., Ayyampalayam S., Burnette J.M., De Paoli E., Estill J.C., Farrell N.P., Harkess A., Jiao Y., Liu K., Mei W.B., Meyers B.C., Shahid S., Wafula E., Zhai J.X., Zhang X.B., Carretero-Paulet L., Lyons E., Tang H.B., Zheng C.F., Altman N.S., Chen F., Chen J.Q., Chiang V., Fogliani B., Guo C.C., Harholt J., Job C., Job D., Kim S., Kong H.Z., Li G.L., Li L., Liu J., Park J., Qi X.S., Rajjou L., Burtet-Sarramegna V., Sederoff R., Sun Y.H., Ulvskov P., Villegente M., Xue J.Y., Yeh T.F., Yu X.X., Acosta J.J., Bruenn R.A., de Kochko A., Herrera-Estrella L.R., Ibarra-Laclette E., Kirst M., Pissis S.P., Poncet V. 2013. The *Amborella* genome and the evolution of flowering plants. Science 342:1241089.

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Barkman T.J., Chenery G., McNeal J.R., Lyons-Weiler J., Ellisens W.J., Moore G., Wolfe A.D., dePamphilis C.W. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc. Natl Acad. Sci. U. S. A. 97:13166–13171.

Bell C.D., Soltis D.E., Soltis P.S. 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97:1296–1303.

Bertioli D., Moretzsohn M., Madsen L., Sandal N., Leal-Bertioli S., Guimaraes P., Hougaard B., Fredslund J., Schauser L., Nielsen A., Sato S., Tabata S., Cannon S., Stougaard J. 2009. An analysis of synteny of *Arachis* with *Lotus* and *Medicago* sheds new light on the structure, stability and evolution of legume genomes. BMC Genomics 10:45.

Betancur R., Li C., Munroe T.A., Ballesteros J.A., Orti G. 2013. Addressing gene-tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62:763–785.

Boussau B., Gouy M. 2006. Efficient likelihood computations with nonreversible models of evolution. Syst. Biol. 55:756–768.

Bowers J.E., Chapman B.A., Rong J.K., Paterson A.H. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438.

Bremer B., Bremer K., Chase M.W., Fay M.F., Reveal J.L., Soltis D.E., Soltis P.S., Stevens P.F., Anderberg A.A., Moore M.J., Olmstead R.G., Rudall P.J., Sytsma K.J., Tank D.C., Wurdack K., Xiang J.Q.Y., Zmarzty S. 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. 161:105–121.

Brinkmann H., Philippe H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Brinkmann H., Van der Giezen M., Zhou Y., De Raucourt G.P., Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.

Burleigh J.G., Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

Capella-Gutiérrez S., Silla-Martínez J.M., Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Conant G.C., Lewis P.O. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol. Biol. Evol. 18:1024–1033.

Cummins C.A., McInerney J.O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60:833–844.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Doyle J.A. 2012. Molecular and fossil evidence on the origin of angiosperms. Annu. Rev. Earth Planet. Sci. 40:301–326.

Doyle J.A., Endress P.K. 2014. Integrating Early Cretaceous fossils into the phylogeny of living angiosperms: ANITA lines and relatives of Chloranthaceae. Int. J. Plant Sci. 175:555–600.

Drew B.T., Ruhfel B.R., Smith S.A., Moore M.J., Briggs B.G., Gitzendanner M.A., Soltis P.S., Soltis D.E. 2014. Another look at the root of the angiosperms reveals a familiar tale. Syst. Biol. 63:368–382.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sorensen M.V., Haddock S.H.D., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

Duvick J., Fu A., Muppirala U., Sabharwal M., Wilkerson M.D., Lawrence C.J., Lushbough C., Brendel V. 2008. PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res. 36:D959–D965.

Edgar R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Enright A.J., van Dongen S., Ouzounis C.A. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30:1575–1584.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Finet C., Timme R.E., Delwiche C.F., Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr. Biol. 20:2217–2222.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Friedman W.E. 2006. Embryological evidence for developmental lability during early angiosperm evolution. Nature 441:337–340.

Friedman W.E., Ryerson K.C. 2009. Reconstructing the ancestral female gametophyte of angiosperms: insights from *Amborella* and other ancient lineages of flowering plants. Am. J. Bot. 96:129–143.

Galtier N., Gouy M. 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–879.

Goremykin V., Nikiforova S., Bininda-Emonds O. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71:319–331.

Goremykin V.V., Nikiforova S.V., Biggs P.J., Zhong B.J., Delange P., Martin W., Woetzel S., Atherton R.A., McLenachan P.A., Lockhart P.J. 2013. The evolutionary root of flowering plants. Syst. Biol. 62:50–61.

Goremykin V.V., Viola R., Hellwig F.H. 2009. Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J. Mol. Evol. 68:197–204.

Gribaldo S., Philippe H. 2002. Ancient phylogenetic relationships. Theor. Popul. Biol. 61:391–408.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hejnol A., Obst M., Stamatakis A., Ott M., Rouse G.W., Edgecombe G.D., Martinez P., Baguñà J., Bailly X., Jondelius U., Wiens M., Müller W.E.G., Seaver E., Wheeler W.C., Martindale M.Q., Giribet G., Dunn C.W. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc. R. Soc. B 276:4261–4270.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hirt R.P., Logsdon J.M., Healy B., Dorey M.W., Doolittle W.F., Embley T.M. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl Acad. Sci. U. S. A. 96:580–585.

Huang H., Knowles L.L. 2009. What is the danger of the anomaly zone for empirical phylogenetics? Syst. Biol. 58:527–536.

Huelsenbeck J.P., Bull J.J., Cunningham C.W. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

Jansen R.K., Cai Z., Raubeson L.A., Daniell H., dePamphilis C.W., Leebens-Mack J., Muller K.F., Guisinger-Bellian M., Haberle R.C., Hansen A.K., Chumley T.W., Lee S.-B., Peery R., McNeal J.R., Kuehl J.V., Boore J.L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl Acad. Sci. U. S. A. 104:19369–19374.

Jermiin L.S., Ho S.Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.

Jiao Y., Leebens-Mack J., Ayyampalayam S., Bowers J., McKain M., McNeal J., Rolf M., Ruzicka D., Wafula E., Wickett N., Wu X., Zhang Y., Wang J., Zhang Y., Carpenter E., Deyholos M., Kutchan T., Chanderbali A., Soltis P., Stevenson D., McCombie R., Pires J., Wong G., Soltis D., dePamphilis C. 2012. A genome triplication associated with early diversification of the core eudicots. Genome Biol. 13:R3.

Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali A.S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N., editor. Mammalian protein metabolism. New York (NY): Academic Press. p. 21–132.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

Lartillot N., Rodrigue N., Stubbs D., Richer J. 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. Syst. Biol. 62:611–615.
Lee E.K., Cibrian-Jaramillo A., Kolokotronis S.O., Katari M.S., Stamatakis A., Ott M., Chiu J.C., Little D.P., Stevenson D.W., McCombie W.R., Martienssen R.A., Coruzzi G., DeSalle R. 2011. A functional phylogenomic view of the seed plants. PLoS Genet. 7:e1002411.
Leebens-Mack J., Raubeson L.A., Cui L.Y., Kuehl J.V., Fourcade M.H., Chumley T.W., Boore J.L., Jansen R.K., 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.
Liu L., Edwards S.V. 2009. Phylogenetic analysis in the anomaly zone. Syst. Biol. 58:452–460.
Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.
Liu L., Yu L. 2010. Phybase: an R package for species tree analysis. Bioinformatics 26:962–963.
Liu L., Yu L., Edwards S.V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10:302.
Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S.V. 2009a. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.
Liu L., Yu L., Pearl D.K., Edwards S.V. 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58:468–477.
Liu Q.P., Xue Q.Z. 2005. Comparative studies on codon usage pattern of chloroplasts and their host nuclear genes in four plant species. J. Genet. 84:55–62.
Magallón S., Castillo A. 2009. Angiosperm diversification through time. Am. J. Bot. 96:349–365.
Mathews S., Donoghue M.J. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286:947–950.
Moore M.J., Bell C.D., Soltis P.S., Soltis D.E. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl Acad. Sci. U. S. A. 104:19363–19368.
Moore M.J., Hassan N., Gitzendanner M.A., Bruenn R.A., Croley M., Vandeventer A., Horn J.W., Dhingra A., Brockington S.F., Latvis M., Ramdial J., Alexandre R., Piedrahita A., Xi Z., Davis C.C., Soltis P.S., Soltis D.E. 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. Int. J. Plant Sci. 172:541–558.
Moore M.J., Soltis P.S., Bell C.D., Burleigh J.G., Soltis D.E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl Acad. Sci. U. S. A. 107:4623–4628.
Nesnidal M.P., Helmkampf M., Bruchhaus I., Hausdorf B. 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol. Biol. Evol. 27:2095–2104.
Olsen G.J. 1987. Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harb. Symp. Quant. Biol. 52:825–837.
Parkinson C.L., Adams K.L., Palmer J.D. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr. Biol. 9:1485–1488.
Pfeil B.E., Schlueter J.A., Shoemaker R.C., Doyle J.J. 2005. Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families. Syst. Biol. 54:441–454.
Philippe H., Roure B. 2011. Difficult phylogenetic questions: more data, maybe; better methods, certainly. BMC Biol. 9:91.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Worheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Philippe H., Lopez P., Brinkmann H., Budin K., Germot A., Laurent J., Moreira D., Müller M., Le Guyader H. 2000. Early-branching or fast-evolving eukaryotes? An answer based on slowly evolving positions. Proc. R. Soc. B 267:1213–1221.
Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the arthropoda. Syst. Biol. 53:978–989.
Qiu Y.L., Dombrovska O., Lee J., Li L.B., Whitlock B.A., Bernasconi-Quadroni F., Rest J.S., Davis C.C., Borsch T., Hilu K.W., Renner S.S., Soltis D.E., Soltis P.S., Zanis M.J., Cannone J.J., Gutell R.R., Powell M., Savolainen V., Chatrou L.W., Chase M.W. 2005. Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. Int. J. Plant Sci. 166:815–842.
Qiu Y.L., Lee J., Bernasconi-Quadroni F., Soltis D.E., Soltis P.S., Zanis M., Zimmer E.A., Chen Z., Savolainen V., Chase M.W. 2000. Phylogeny of basal angiosperms: analyses of five genes from three genomes. Int. J. Plant Sci. 161:S3–S27.
Qiu Y.L., Lee J., Bernasconi-Quadroni F., Soltis D.E., Soltis P.S., Zanis M., Zimmer E.A., Chen Z.D., Savolainen V., Chase M.W. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404–407.
Qiu Y.L., Li L., Wang B., Xue J.-Y., Hendry T.A., Li R.-Q., Brown J.W., Liu Y., Hudson G.T., Chen Z.-D. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J. Syst. Evol. 48:391–425.
Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Rannala B., Yang Z. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164:1645–1656.
Rosenberg N.A., Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57:131–140.
Salemi M., Vandamme A.-M. 2003. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge, UK: Cambridge University Press.
Seo T.K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.
Sheffield N.C., Song H.J., Cameron S.L., Whiting M.F. 2009. Nonstationary evolution and compositional heterogeneity in beetle mitochondrial phylogenomics. Syst. Biol. 58:381–394.
Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.
Shimodaira H. 2008. Testing regions with nonsmooth boundaries via multiscale bootstrap. J. Stat. Plan. Infer. 138:1227–1241.
Shulaev V., Sargent D.J., Crowhurst R.N., Mockler T.C., Folkerts O., Delcher A.L., Jaiswal P., Mockaitis K., Liston A., Mane S.P., Burns P., Davis T.M., Slovin J.P., Bassil N., Hellens R.P., Evans C., Harkins T., Kodira C., Desany B., Crasta O.R., Jensen R.V., Allan A.C., Michael T.P., Setubal J.C., Celton J.-M., Rees D.J.G., Williams K.P., Holt S.H., Rojas J.J.R., Chatterjee M., Liu B., Silva H., Meisel L., Adato A., Filichkin S.A., Troggio M., Viola R., Ashman T.-L., Wang H., Dharmawardhana P., Elser J., Raja R., Priest H.D., Bryant D.W., Fox S.E., Givan S.A., Wilhelm L.J., Naithani S., Christoffels A., Salama D.Y., Carter J., Girona E.L., Zdepski A., Wang W., Kerstetter R.A., Schwab W., Korban S.S., Davik J., Monfort A., Denoyes-Rothan B., Arus P., Mittler R., Flinn B., Aharoni A., Bennetzen J.L., Salzberg S.L., Dickerman A.W., Velasco R., Borodovsky M., Veilleux R.E., Folta K.M. 2011. The genome of woodland strawberry (*Fragaria vesca*). Nat. Genet. 43:109–116.
Smith S.A., Dunn C.W. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716.
Smith S.A., Beaulieu J.M., Donoghue M.J. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc. Natl Acad. Sci. U. S. A. 107:5897–5902.
Soltis D.E., Soltis P.S. 2004. *Amborella* not a "basal angiosperm"? Not so fast. Am. J. Bot. 91:997–1001.
Soltis D.E., Bell C.D., Kim S., Soltis P.S. 2008. Origin and early evolution of angiosperms. Ann. N. Y. Acad. Sci. 1133:3–25.
Soltis D.E., Gitzendanner M.A., Soltis P.S. 2007. A 567-taxon data set for angiosperms: the challenges posed by Bayesian analyses of large data sets. Int. J. Plant Sci. 168:137–157.
Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D.,

Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98:704–730.

Soltis D.E., Soltis P.S., Chase M.W., Mort M.E., Albach D.C., Zanis M., Savolainen V., Hahn W.H., Hoot S.B., Fay M.F., Axtell M., Swensen S.M., Prince L.M., Kress W.J., Nixon K.C., Farris J.S. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Bot. J. Linn. Soc. 133:381–461.

Soltis P.S., Soltis D.E., Chase M.W. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402–404.

Song S., Liu L., Edwards S.V., Wu S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl Acad. Sci. U. S. A. 109: 14942–14947.

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

Stefanović S., Rice D.W., Palmer J.D. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol. 4:35.

Suyama M., Torrents D., Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612.

Tang H., Bowers J.E., Wang X., Paterson A.H. 2010. Angiosperm genome comparisons reveal early polyploidy in the monocot lineage. Proc. Natl Acad. Sci. U. S. A. 107:472–477.

Wang H., Moore M.J., Soltis P.S., Bell C.D., Brockington S.F., Alexandre R., Davis C.C., Latvis M., Manchester S.R., Soltis D.E. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc. Natl Acad. Sci. U. S. A. 106:3853–3858.

Wasmuth J.D., Blaxter M.L. 2004. prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics 5:187.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Wikström N., Savolainen V., Chase M.W. 2001. Evolution of the angiosperms: calibrating the family tree. Proc. R. Soc. B 268:2211–2220.

Wodniok S., Brinkmann H., Glockner G., Heidel A., Philippe H., Melkonian M., Becker B. 2011. Origin of land plants: do conjugating green algae hold the key? BMC Evol. Biol. 11:104.

Wu Y. 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66:763–775.

Xi Z., Rest J.S., Davis C.C. 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. PLoS One 8:e80870.

Xia X., Xie Z. 2001. DAMBE: software package for data analysis in molecular biology and evolution. J. Hered. 92:371–373.

Xia X., Xie Z., Salemi M., Chen L., Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26:1–7.

Zanis M.J., Soltis D.E., Soltis P.S., Mathews S., Donoghue M.J. 2002. The root of the angiosperms revisited. Proc. Natl Acad. Sci. U. S. A. 99:6848–6853.

Zhang N., Zeng L., Shan H., Ma H. 2012. Highly conserved low-copy nuclear genes as effective markers for phylogenetic analyses in angiosperms. New Phytol. 195:923–937.

Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3:1340–1348.

Zhong B., Liu L., Yan Z., Penny D. 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18:492–495.