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

ZHENXIANG XI1, LIANG LIU2, JOSHUA S. REST3, AND CHARLES C. DAVIS1,*

1Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA; 2Department of Statistics and Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA; and 3Department 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

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., Nymphaeae) 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 Nymphaeae 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 (Huaelsenbeck 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 lineage of extant angiosperms. By incorporating transcriptomes to investigate the earliest diverging genome sequencing projects and deeply sequenced has not been more thoroughly investigated.

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 (Pinus 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 (Pinus morrisonicola Hayata, Pinus koraiensis Siebold & Zucc., and Cedrus 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} \text{-value threshold, and then grouped with MCL v9.0-308 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-Gutierrez 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 GTR+GAMMA 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 Dunne 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
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 GTR+GAMMA 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 ($I_{SS}$) as implemented in DAMBE (Xia and Xie 2003) or nhPhyML (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; Supplementary Fig. S1). For each nuclear gene, the branch lengths of the assigned topology and parameters of the GTR+GAMMA model were estimated from the original nucleotide sequences using RAxML with the “-d 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 + $\Gamma$ 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,
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 each species tree, 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;
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 $\mu$ is the average mutation rate per site per generation (Liu and Yu 2010). We applied two different values of $\theta$ to simulate varying degrees of deep coalescence (i.e., $\theta = 0.0001$ for the species trees 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 $\theta$, and a large value of $\theta$ 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 $\theta$. Therefore, the species tree 6T-1 produced gene trees with highly discordant topologies (i.e., a
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 (442,596 parsimony informative sites), 279% 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), 51.1% 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 + magnolids, (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 magnolids are sister to eudicots + monocots, (ii) Solanales are sister to Gentianales + Lamiales, (iii) Myrtales are sister to fabids + 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 identify 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 (Wikstrom
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. S5) and slow plastid gene (Supplementary Fig. S6) 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
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) 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).

### Gene trees (n = 310)

<table>
<thead>
<tr>
<th>Gene trees (n = 310)</th>
<th>OV - slow</th>
<th>OV - fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>STAR</td>
<td>MP-EST</td>
<td>RAxML</td>
</tr>
<tr>
<td>0%:100%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>10%:90%</td>
<td>1.0</td>
<td>0.99</td>
</tr>
<tr>
<td>20%:80%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>30%:70%</td>
<td>1.0</td>
<td>0.97</td>
</tr>
<tr>
<td>40%:60%</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>50%:50%</td>
<td>0.66</td>
<td>0.60</td>
</tr>
<tr>
<td>60%:40%</td>
<td>1.0</td>
<td>0.95</td>
</tr>
<tr>
<td>70%:30%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>80%:20%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>90%:10%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>100%:0%</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

**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).
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 concatenation 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 from 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, 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% (Aquilégia coerulea James) to 53.7% (Selaginella) for nuclear genes (Supplementary Table S1) and from 32.2% (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; Bouckaert and Gouy 2006) as implemented in nPhyML. 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 (Solits and Solits 2004; Stefanovic et al. 2004), we re-analyzed the recent 649-species plastid gene data set from Solits 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 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 Roux 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 Etherid, 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**


Page: 350 313-322

[15:18 3/10/2014 Sysbio-syu551.txt]


Soltis S.E., Smith S.A., Collinse N., Wyeraider K.J., Tank D.C., Brockington S.F., Refullo-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawford S., Black C., Doud D,


