Expanded Taxonomic Sampling Coupled with Gene Genealogy Interrogation Provides Unambiguous Resolution for the Evolutionary Root of Angiosperms

Bojian Zhong[1,*,†] and Ricardo Betancur-R[2,*,†]

[1]College of Life Sciences, Nanjing Normal University, China

[2]Department of Biology, University of Puerto Rico – Río Piedras, San Juan, Puerto Rico

†These authors contributed equally to this work.

*Corresponding authors: E-mails: bjzhong@gmail.com; betanri@fishphylogeny.org.

Accepted: November 7, 2017

Abstract

The branching order of major angiosperm lineages is a challenging phylogenetic question that has received substantial attention in recent years. Two main competing hypotheses place the New Caledonian *Amborella* as either sister to all other extant angiosperms (*Amborella*-sister) or to the water lilies (*Amborella* + Nymphaeales). Here, we revisit this question by expanding a transcriptomic data set of 310 genes previously assembled to include data from seven species comprising two major lineages of flowering plants that were poorly represented or missing from the original study. We also applied gene genealogy interrogation, a recent approach based on constrained tree searches in combination with topology tests, to account for gene tree estimation error and its downstream effects in coalescent analyses. In addition to gene genealogy interrogation, we conducted a large number of multilocus analyses, including concatenation and coalescent approaches (using both unconstrained and constrained gene trees), and based on different data sets (original and expanded) and data types (nucleotide and amino acid sequences). We show that the majority of gene trees favor *Amborella*-sister topology, and all multilocus analyses conducted (concatenation and coalescent) provide overwhelming support for this hypothesis regardless of data type. Beyond resolving the evolutionary root of angiosperms with confidence, our results highlight the importance of both broadening taxonomic sampling in phylogenomics and addressing the effects of gene tree error in summary coalescent inferences.

Key words: root of angiosperms, *Amborella*, taxon sampling, gene genealogy interrogation, gene tree estimation error.

With nearly 300,000 species, angiosperms (flowering plants) comprise the largest clade of terrestrial plants (Judd et al. 1999). Although important efforts have been made to resolve the phylogeny of angiosperms (e.g., Soltis et al. 2011; APG IV 2016), the branching order of major lineages at the root remains a challenging phylogenetic problem. There are currently two competing hypotheses concerning the extant sister group to all other flowering plants: 1) *Amborella trichopoda* alone—a unique shrub species endemic to New Caledonia in the order Amborellales (e.g., Mathews and Donoghue 1999; Soltis et al. 1999, 2000; Stefanović et al. 2004; Leebens-Mack et al. 2005; Jansen et al. 2007; Graham and Iles 2009; Moore et al. 2011; Soltis et al. 2011; Drew et al. 2014; Ruhfel et al. 2014; Wickett et al. 2014; Zeng et al. 2014; Simmons and Gatesy 2015; Simmons 2016) or 2) a clade comprising *Amborella* plus the water lilies (Nymphaeales) (e.g., Barkman et al. 2000; Soltis et al. 2000; Leebens-Mack et al. 2005; Jansen et al. 2006; Moore et al. 2007; Graham and Iles 2009; Qiu et al. 2010; Finet et al. 2012; Laurin-Lemay et al. 2012; Goremykin et al. 2013; Xi et al. 2014). A third possibility involves placement of water lilies as sister to all other angiosperms, but this topology has received little support from molecular phylogenetic analyses (but see Yang et al. 2007; fig. 1).

To address the relationships of major angiosperm lineages, previous studies have assembled molecular data sets including a few to hundreds of genes (e.g., Soltis et al. 2000; Jansen et al. 2007; Moore et al. 2007; Wang et al. 2009; Lee et al. 2011; Moore et al. 2011; Soltis et al. 2011; Drew et al. 2014; Xi et al. 2014; Zeng et al. 2014). Prior to 2014, all analyses conducted were based on concatenated alignments of all gene fragments examined. Both simulation and empirical data have shown, however, that concatenation approaches

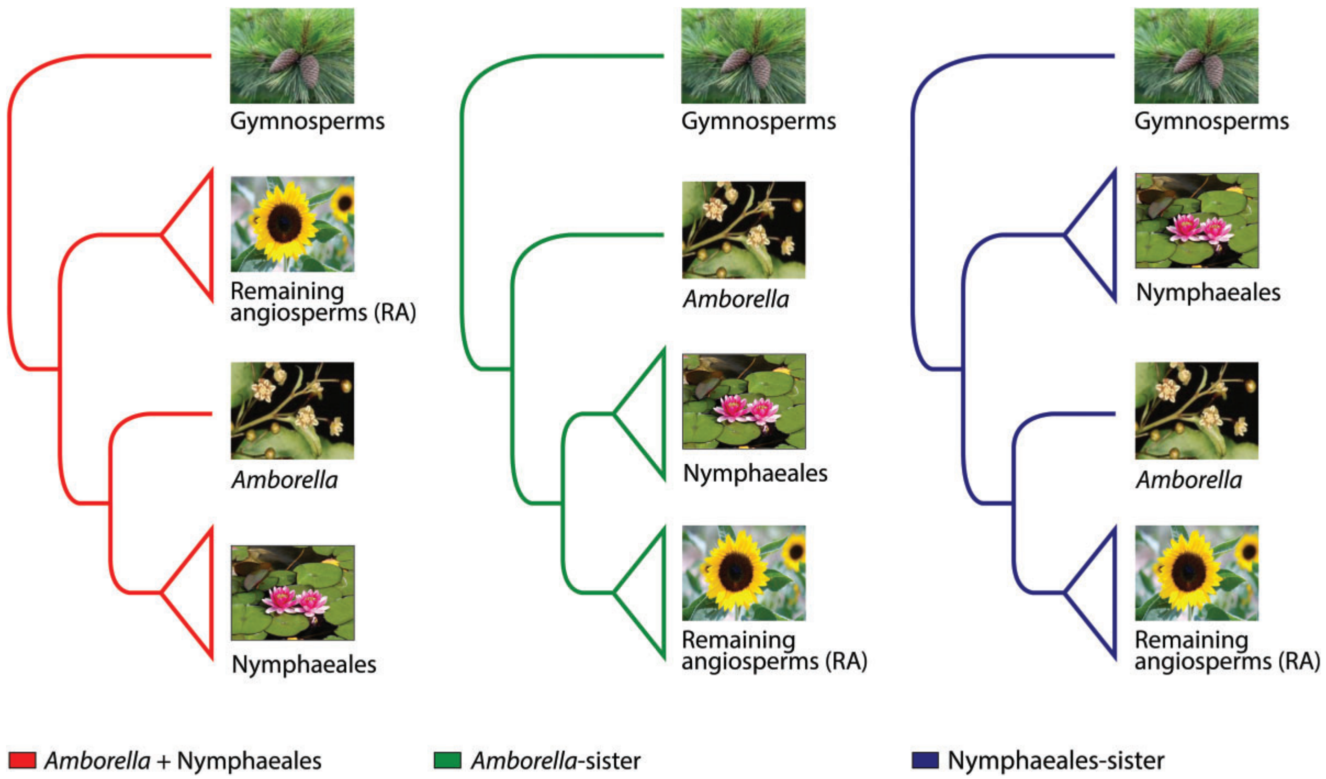

Fig. 1.—Three possible topologies depicting the branching order of three major angiosperm lineages (*Amborella*, Nymphaeales, and the RA clade). See text (first paragraph) for a list of citations supporting each tree.

can yield misleading results when the underlying assumption of nondiscordant gene histories is extensively violated by the data (Kubatko and Degnan 2007; Liu and Edwards 2009). Recent development of coalescent methods that can better accommodate gene tree heterogeneity resulting from the presence of incomplete lineage sorting (ILS) provides an avenue for species tree inference (Liu 2008; Kubatko et al. 2009; Liu et al. 2009; Heled and Drummond 2010; Larget et al. 2010; Liu et al. 2010; Mirarab et al. 2014; Mirarab and Warnow 2015).

Xi et al. (2014) applied both concatenation and coalescent approaches to investigate the evolutionary root of angiosperms using a transcriptomic data set consisting of 310 nuclear genes and 46 taxa, including most major angiosperm lineages as well as gymnosperm outgroups. Although their concatenation analyses provide support for the *Amborella*-sister hypothesis, summary coalescent approaches (MP-EST and STAR; Liu et al. 2009, 2010) applied to this data set resolved *Amborella* + Nymphaeales tree. Xi et al. (2014) dismissed the results based on concatenation methods invoking their inability to account for ILS. These authors further reinforced their conclusions following the results obtained using tree-independent character-subsampling procedures to filter out fast-evolving sites (see also Edwards et al. 2016). Recent reanalyses of Xi et al. (2014)'s data set, however, have shown that the summary coalescent approaches used are sensitive to both mis-rooted gene trees and alignment errors (Simmons and Gatesy 2015), and that the character subsampling strategies utilized are biased (Simmons 2016). When those two factors are accounted for, *Amborella*-sister is resolved even with coalescent approaches (Simmons and Gatesy 2015; Simmons 2016).

Two main methodological artifacts compromising the accuracy of phylogenetic analyses have not yet been considered by the recent genomic studies that investigate the evolutionary origins of major angiosperm lineages: 1) gene tree estimation error (i.e., inferred gene trees failing to depict the true genealogical history of genes) and 2) detrimental effects of limited taxonomic sampling. The former factor typically stems from the low information content of the (often short) individual gene fragments (e.g., Roch and Warnow 2015; Arcila et al. 2017), whereas failure to include key lineages that bisect long branches can produce systematic errors such as long-branch attraction (e.g., Wiens 2003; Heath et al. 2008). Aside from the mis-rooting issues raised (Simmons and Gatesy 2015; Simmons 2016), gene tree error was discussed but not directly addressed by the previous studies. Furthermore, the transcriptomic data set analyzed by Xi et al. (2014) included a single species of water lilies and lacked early-branching representatives in the "remaining angiosperms" unnamed clade (RA clade hereafter).

Here, we revisit the root of angiosperms problem by accounting for potential systematic biases arising from limited taxon sampling and gene tree estimation error. We expanded

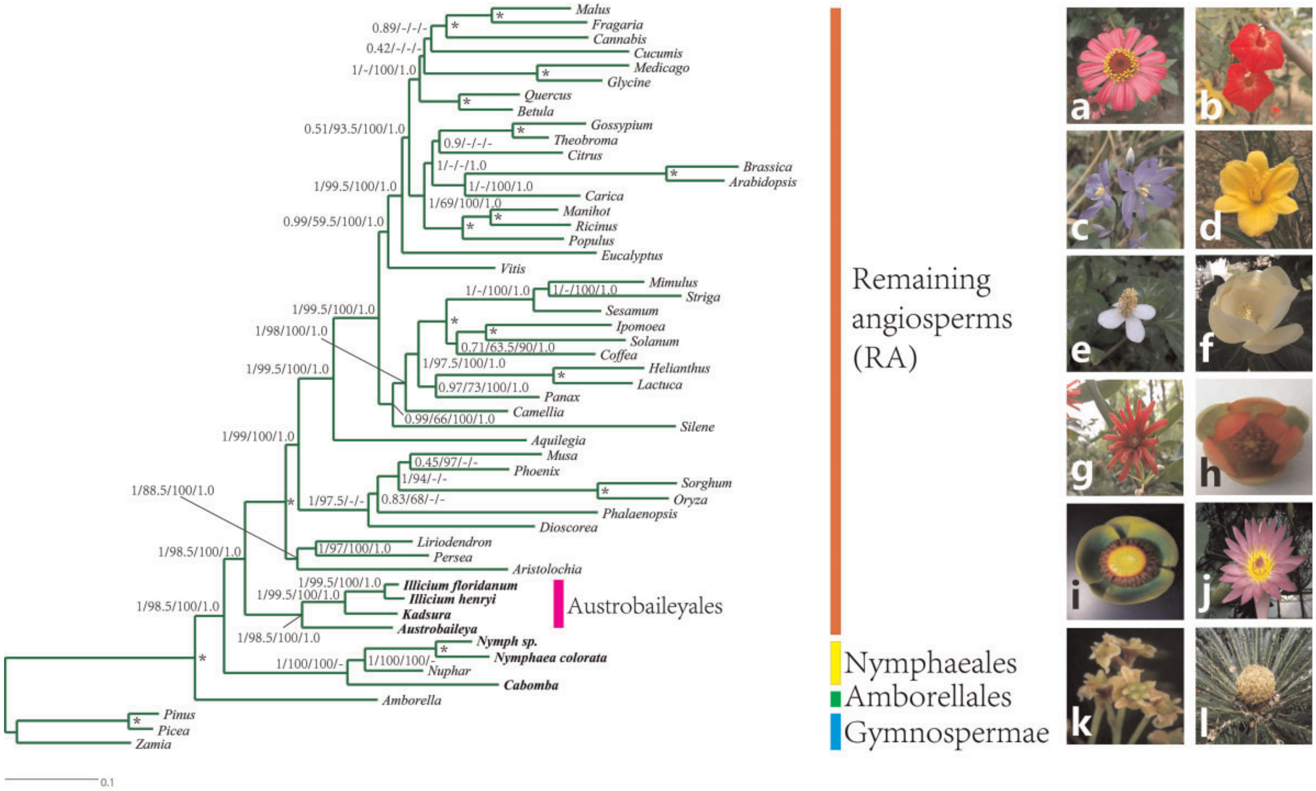

Fig. 2.—Selected tree based on the ASTRAL-II analysis of the expanded data set using nucleotide sequences. Branch lengths shown were estimated for the concatenated matrix using RAxML. Nodal support values are indicated above each branch (posterior probability or bootstrap support obtained from ASTRAL-II/MP-EST/RAxML/ExaBayes); an asterisk indicates the clade is fully supported. New taxa added here are shown in bold. All other trees derived from coalescent (4 other analyses), concatenation (2 analyses), and GGI-based coalescent analyses (32 analyses) also resolve the *Amborella*-sister tree (provided as Supplementary Material online; see support values in table 1). Photographs are examples of seed plant diversity: (*a*) *Zinnia elegans*, (*b*) *Ipomoea* × *sloteri*, (*c*) *Monochoria korsakowii*, (*d*) *Hemerocallis fulva*, (*e*) *Houttuynia cordata*, (*f*) *Magnolia grandiflora*, (*g*) *Illicium floridanum*, (*h*) *Schisandra sphenanthera*, (*i*) *Nuphar advena*, (*j*) *Nymphaea tetragona*, (*k*) *Amborella trichopoda*, (*l*) *Cycas pectinata*.

Xi et al. (2014)'s transcriptomic data set of 310 nuclear genes to include previously unexamined sequences from two early-diverging angiosperm lineages. The new taxa added include three water lilies that split the long branch leading to *Nuphar* (sole representative of Nymphaeales in Xi et al. 2014's data set) and four species in Austrobaileyales, which comprises the earliest-branching lineage in the RA clade (not represented in the original data set; see APG IV 2016; fig. 2).

We also conducted gene genealogy interrogation (GGI), a recently proposed approach to discern between estimation error and actual biological conflict (e.g., resulting from ILS) in explaining gene tree incongruence (Arcila et al. 2017). This method interrogates individual genes via topology tests to determine the genealogical history that each gene supports with highest probability. In other words, it identifies which tree topology, from a set of predefined topologies, provides the best fit for each of the (often short) gene partitions. GGI extracts the signal from genes by applying topological constraints according to the number of alternative trees for a given *n*-taxon statement. The base of angiosperms is a rooted 3-taxon (or unrooted 4-taxon) problem that involves three possible topologies: *Amborella*-sister, *Amborella* + Nymphaeales, or Nymphaeales-sister (see fig. 1). The resulting constrained trees for each gene are statistically compared and ranked according to the *P* values obtained using the approximately unbiased (AU) test (Shimodaira 2002). In addition to summarizing the relative support that each alternative topology receives across genes, the resulting rank 1 gene trees selected by the topology tests (GGI gene trees) can be used as input for species tree methods, thereby accounting for systematic errors intrinsic to unconstrained gene tree inference and ultimately summary coalescent analyses (Arcila et al. 2017; Mirarab 2017).

To assemble the expanded data set, we first excluded the highly divergent outgroup *Selaginella moellendorffii* (included by Xi et al. 2014) to avoid systematic biases stemming from long-branch attraction (see Simmons and Gatesy 2015). Some analyses (GGI; see below), however, used the original and expanded data sets both with and without *Selaginella* for more direct comparison with previous studies. All analyses conducted were based on both nucleotide and amino acid sequences. We first inferred unconstrained gene trees under

Table 1

Nodal Support Values for the RA + Nymphaeales Clade (to the Exclusion of *Amborella*) Obtained from Different Analyses and Data Set Types

Method	Data Set Type	Model	Nodal Support	Figure
RAxML (concatenation)	Nucleotide	GTRGAMMA	100 (BS)	S4
ExaBayes (concatenation)	Nucleotide	GTRGAMMA	1.00 (PP)	S5
ASTRAL-II (coalescent)	Nucleotide	GTRGAMMA	1.00 (PP)	S6
ASTRAL-II (coalescent)	Nucleotide	GTRGAMMA	98.2 (BS)	S7
MP-EST (coalescent)	Nucleotide	GTRGAMMA	98.5 (BS)	S8
ASTRAL-II (coalescent)	Nucleotide	CATGTR	1.00 (PP)	S9
RAxML (concatenation)	Amino acid	PROTGAMMAWAG	100 (BS)	S10
ExaBayes (concatenation)	Amino acid	PROTGAMMAWAG	1.00 (PP)	S11
IQ-TREE (concatenation)	Amino acid	LG+C20+F+G	100 (BS)	S12
ASTRAL-II (coalescent)	Amino acid	PROTGAMMAWAG	1.00 (PP)	S13
ASTRAL-II (coalescent)	Amino acid	PROTGAMMAWAG	94.6 (BS)	S14
MP-EST (coalescent)	Amino acid	PROTGAMMAWAG	89.5 (BS)	S15
ASTRAL-II (coalescent)	Amino acid	CATGTR	1.00 (PP)	S16
ASTRAL-II (coalescent)	Amino acid	LG+C20+F+G	0.99 (PP)	S17
PhyloBayes (concatenation)[a]	Nucleotide	CATGTR	1.00 (PP)	S18
ASTRAL-II (coalescent)[a]	Nucleotide	GTRGAMMA	1.00 (PP)	S19
MP-EST (coalescent)[a]	Nucleotide	GTRGAMMA	100 (BS)	S20

Note.—The model shown under coalescent methods is the substitution model used for gene tree estimation. The unconstrained gene trees were used for coalescent methods. BS, bootstrap support; PP, posterior probability.

[a] Analyses that were conducted using the top 100 most conserved genes.

maximum likelihood (ML) in RAxML and IQ-TREE, and Bayesian inference in PhyloBayes using infinite mixtures. The resulting trees were input for summary species tree analyses under two different approaches: ASTRAL-II and MP-EST. We then conducted concatenation analyses under ML (RAxML and IQ-TREE) and Bayesian (ExaBayes and PhyloBayes) criteria using the expanded data sets. To account for biases arising from the inclusion of fast-evolving sites (Xi et al. 2014; Edwards et al. 2016), we also selected the top 100 most conserved genes for both concatenation (PhyloBayes) and species tree (ASTRAL and MP-EST) inference.

Next, we applied GGI to test the three alternative topologies concerning the evolutionary root of angiosperms (see above). We conducted a total of 32 GGI analyses, comprising different data sets (original and expanded, both with and without *Selaginella*), data types (nucleotide and amino acid sequences), using two alternative approaches for sampling GGI trees (complete data set including all rank 1 trees and a subset of all rank 1 trees that are significantly better than the alternatives), and applying different numbers of constraints. Following points raised by Mirarab (2017), alternative scenarios for topological constraints include enforcing (as originally proposed) or relaxing the monophyly of the RA and Nymphaeales subclades (see supplementary fig. S1, Supplementary Material online). By relaxing the assumption of subclade monophyly, we account for the potential of ILS to disrupt subclades in a nontrivial proportion of genes. Finally, we used 16 (of 32) sets of GGI gene trees, obtained with two of the most contrasting data sets assembled (the original including *Selaginella* and the expanded excluding it), as input for ASTRAL-II and MP-EST (32 analyses). See Materials and Methods for additional details.

Regardless of the method (coalescent or concatenation) or data type (nucleotide or amino acids) used, all multilocus analyses conducted support the *Amborella*-sister topology (fig. 2 and supplementary figs. S4–S20, Supplementary Material online). Most concatenation analyses (RAxML, IQ-TREE, ExaBayes, and PhyloBayes) and coalescent-based ASTRAL-II (with different data types, substitution models, and methods applied for gene tree estimation; see details in table 1) resolved a clade including RA + Nymphaeales (to the exclusion of *Amborella*) with full support (table 1). The same is true for concatenation and coalescent analyses based on a subset of the most conserved genes (table 1). Although multilocus bootstrapping support obtained with ASTRAL-II and MP-EST is not as strong (89.5–98.5%; table 1), both coalescent methods also resolved this clade.

All GGI analyses, based on both the original and expanded data sets, resulted in most genes favoring *Amborella*-sister (mean frequency across analyses based on all GGI gene trees or $\bar{x}_{all} = 44.7\%$; mean frequency across analyses based on significant-only GGI gene trees or $\bar{x}_{sign.} = 66.8\%$) over *Amborella* + Nymphaeales ($\bar{x}_{all} = 28.0\%$; $\bar{x}_{sign.} = 20.3\%$) and Nymphaeales-sister ($\bar{x}_{all} = 27.4\%$; $\bar{x}_{sign.} = 12.9\%$; fig. 3 and supplementary figs. S2 and S3, Supplementary Material online). The 32 coalescent-based analyses that used the GGI gene trees as input also resolved this tree (MP-EST and ASTRAL-II; available from FigShare). Although we did not conduct a modified version of GGI that uses constrained topologies in combination with statistically better unconstrained

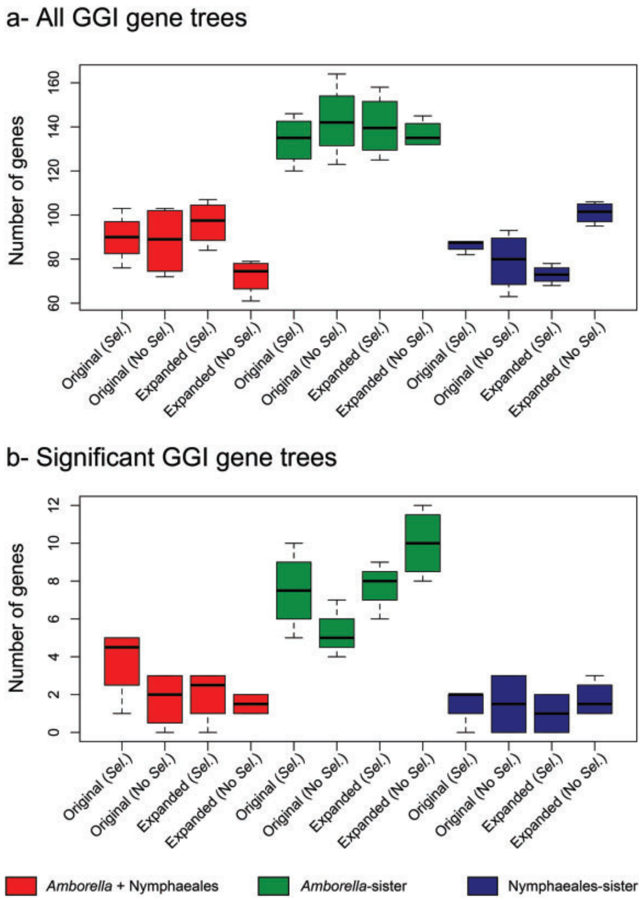

Fig. 3.—Results from GGI applied to the original or expanded data sets, based on either (*a*) all gene trees or (*b*) a subset of rank 1 gene trees with ML searches that are significantly better ($P < 0.05$) than the alternatives. Each boxplot summarizes data points obtained from four alternative analyses: using either nucleotide or amino acid sequences, and enforcing (via topological constraints) or relaxing (unconstraining) the assumption of monophyly for the two main subclades (i.e., Nymphaeales [expanded data set only] and RA). See supplementary figs. S2 and S3, Supplementary Material online, for individual results based on the 16 analyses conducted. *Sel.*: data set including the *Selaginella* outgroup; No *Sel.*: data set excluding it.

gene trees (suggested by Mirarab 2017), we note that species tree analyses using either fully unconstrained (non-GGI) or fully constrained (GGI) tree searches all resolved the *Amborella*-sister tree.

Both the original and expanded data sets use the same set of 310 protein-coding genes; they differ in that the latter excludes a highly divergent gymnosperm outgroup (*S. moellendorffii*) and includes seven additional lineages that branch off near the angiosperm root, bisecting two long branches (Nymphaeales and Austrobaileyales in the RA clade; see fig. 2). Unlike the conflict between concatenation and coalescent analyses reported by Xi et al. (2014) based on the original data set, our multilocus analyses using the expanded data set appear to be robust to method choice. Additionally, our examination of the distribution of GGI gene trees shows that *Amborella*-sister is the most frequent gene genealogy—not only with the expanded data set (with or without *Selaginella*), but perhaps most importantly with the original data set (fig. 3 and supplementary figs. S2 and S3, Supplementary Material online). Inability of the Xi et al. (2014) study to achieve this result using unconstrained tree searches in combination with coalescent analyses suggests that their results may have been affected by gene tree error and/or incomplete taxon sampling.

Although all GGI analyses resulted in broadly similar gene tree distributions, some disparity across data sets exists. For instance, boxplots based on analyses derived from two alternative data sets reveal nonoverlapping frequencies of significant GGI trees in favor of *Amborella*-sister (fig. 3*b*). Likewise, analysis of one of the expanded data sets assembled (all genes, without *Selaginella*) find higher support for Nymphaeales-sister over *Amborella* + Nymphaeales (fig. 3*a*), which is opposite to what other data sets resolve.

Another recent study (Shen et al. 2017) applied a method similar to GGI to address the root of angiosperms (along with other recalcitrant groups in the Tree of Life) based on a single analysis obtained with a different transcriptomic data set (assembled by Wickett et al. 2014). The two methods are similar in that they both estimate the difference in support for alternative trees across genes using topological constraints in combination with the AU test. The major difference is that GGI considers all possible topologies for a given *n*-taxon problem, whereas the other approach compares the top two most contentious trees only. Although the Shen et al. (2017) results also provide higher statistical support for *Amborella*-sister relative to *Amborella* + Nymphaeales, the authors conducted no multilocus analyses (concatenation or coalescent) and failed to consider the Nymphaeales-sister tree, which is favored by a substantial number of genes, with one analysis even resolving it at a higher frequency than *Amborella* + Nymphaeales (fig. 3*a*).

In summary, our investigation of the evolutionary root of flowering plants using GGI in combination with expanded taxonomic sampling of critical lineages consistently resolve the *Amborella*-sister topology as both the organismal phylogeny receiving greatest support (fig. 2; table 1) and the most frequent gene genealogy (fig. 3 and supplementary figs. S2 and S3, Supplementary Material online). Our study ultimately underscores the importance of addressing gene tree error when implementing coalescent approaches and the necessity of broadening taxonomic coverage in phylogenomics. Diverting from "many-genes few-taxa" approaches is a crucial step toward advancing resolution of challenging groups in the Tree of Life.

Materials and Methods

Seven species from Austrobaileyales and Nymphaeales were added to the Xi et al.'s (2014) data set based on in silico mining of genomic or transcriptomic data. The draft genome of *Nymphaea colorata* was kindly provided by L. S. Zhang (unpublished). The assembled transcriptomes of *Cabomba caroliniana* and *Illicium henryi* are from Zeng et al. (2014). The *Austrobaileya scandens*, *Illicium floridanum*, *Kadsura heteroclite*, and *Nymphaea* sp. transcriptomes were retrieved from the "1000 plants" project database (https://db.cngb.org/blast4onekp). The 310 orthologous genes common to the Xi et al.'s (2014) data set were filtered for the newly included taxa using TBlastN searches (Altschul et al. 1990) with an *e*-value 1e−20 and using *Arabidopsis thaliana* as reference. The original data set includes 221.7 genes present per species on average (90–310); the seven species added include 292.4 genes on average (278–303). Thus, missing data is not a factor biasing the properties of the expanded data set relative to the original set (supplementary table S2, Supplementary Material online).

The expanded data set was aligned in MUSCLE (Edgar 2004) on a gene-by-gene basis. The resulting alignments were trimmed in GBlocks v.0.91b (Castresana 2000) using the −*t* = *c* and −*b*5 = half options. Each alignment was visually checked by considering the reading frame and obvious mis-alignment errors were corrected. The corresponding amino acid sequence alignments were generated by translation of the nucleotide alignments in MEGA7 (Kumar et al. 2016). Three gymnosperms outgroups were retained from the original data set (*Zamia*, *Picea*, and *Pinus*) after excluding *S. moellendorffii* (see above).

The 310 gene trees used as input for summary coalescent analyses were inferred in RAxML v8.2 (Stamatakis 2014) using the GTRGAMMA and PROTGAMMAWAG models for nucleotide and amino acid sequence data, respectively. Additionally, gene tree estimation was also conducted under the more complex CATGTR Bayesian model in PhyloBayes (Lartillot et al. 2013) for both nucleotide and amino acid sequences and under the LG+C20+F+G model in IQ-TREE (Nguyen et al. 2015) for amino acid sequences. Two summary coalescent methods were used for species tree inference based on the alternative sets of gene trees obtained. The ASTRAL-II v 4.10.12 (Mirarab et al. 2014; Mirarab and Warnow 2015) analyses used unrooted gene trees with multilocus bootstrapping (Seo 2008) and local posterior probability support (Sayyari and Mirarab 2016). The MP-EST (Liu et al. 2010) analyses used rooted gene trees and multilocus bootstrapping support (24 gene trees were discarded for MP-EST as they lacked outgroups for tree rooting after excluding *S. moellendorffii*). Finally, to address the results by Xi et al. (2014) showing that support for *Amborella* + Nymphaeales increases when fast-evolving sites are filtered out (but see points raised by Simmons 2016 about potential biases in the sampling strategy utilized), we selected the top 100 most conserved genes with highest average pairwise similarity (81–97%; following Betancur-R. et al. 2014) for both coalescent and concatenation analyses.

For concatenation analyses, the nucleotide alignments were partitioned by gene and codon positions; amino acid alignments used by-gene partitions only. Best-fit partitioning schemes were first selected using the Akaike information criterion (AIC) as implemented in Partitonfinder2 (Lanfear et al. 2017). A total of 86 and 49 partitions were selected by the AIC for nucleotide and amino acid alignments, respectively, and applied accordingly for downstream analyses. The ML concatenation trees were estimated using the GTRGAMMA (RAxML) model for nucleotide data and the PROTGAMMAWAG (RAxML) and LG+C20+F+G models (IQ-TREE, Nguyen et al. 2015) for amino acid data. Bayesian phylogenetic inference was conducted in ExaBayes v1.5 (Aberer et al. 2014) using two independent runs, each with 4 chains and run for 1,000,000 generations. Finally, we also conducted Bayesian concatenation analyses under the CAT+GTR mixture model in PhyloBayes. To reduce computational time with PhyloBayes, the data set was filtered to include the top 100 most conserved genes, as explained above. Convergence of chains for Bayesian analyses was assessed by ensuring that average standard deviations of split frequencies were lower than 5%.

The 32 sets of GGI analyses conducted involved a total of 7,440 constrained ML searches: 3 topologies for each of 310 gene trees, based on nucleotide or amino acid sequences, enforcing or relaxing the monophyly of the Nymphaeales (expanded data set only) and RA subclades, and including or excluding the highly divergent *Selaginella* outgroup (see Simmons and Gatesy 2015). Site likelihood scores for each alternative tree were obtained with RAxML, and a topology test was conducted for each gene by statistically comparing the scores of the three trees via the AU test (Shimodaira 2002) as implemented in CONSEL v0.1 (Shimodaira and Hasegawa 2001). Trees were ranked according to the *P* values and visualized using box plots, cumulative plots, and columns charts. Two alternative approaches were then applied for sampling GGI trees, one using all rank 1 trees (complete data with 310 genes) and another using the set of rank 1 trees that are significantly better than the alternatives (smaller subset; fig. 3 and supplementary figs. S2 and S3, Supplementary Material online). Finally, 16 (of 32) sets of GGI trees selected from the previous step, based on 2 of the 4 data sets assembled (see main text), were used as input for ASTRAL-II and MP-EST (32 analyses).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Pamela S. Soltis, Douglas E. Soltis, Cymon J. Cox and two anonymous reviewers for their valuable comments. We also thank Liangsheng Zhang and Hong Ma for providing genomic and transcriptomic data; Lingxiao Yang and Xin Chang for help in preparing data and figures; Ning Zhang and Hong Ma for providing plant photographs. This work was supported by the National Science Foundation (NSF) grants (DEB-147184, DEB-1541491) to R.B.R. and by the National Natural Science Foundation of China (31570219), the Natural Science Foundation of Jiangsu Province (BK20150971), the Jiangsu Province Key Project for Scientific Research (16KJA180002), the Six Talent Peaks Projects of Jiangsu Province (2016-XNY-035), the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), and Qing Lan Project to B.Z.

Literature Cited

Aberer AJ, Kobert K, Stamatakis A. 2014. ExaBayes: massively parallel Bayesian tree inference for the whole-genome era. Mol Biol Evol. 31(10):553–2556.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215(3):729–731.

APG IV [Angiosperm Phylogeny Group IV]. 2016. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc. 181(1):1–20.

Arcila D, et al. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat Ecol Evol. 1(2):0020.

Barkman TJ, et al. 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(24):13166–13171.

Betancur-R R, Naylor G, Orti G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst Biol. 63(2):257–262.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17(4):540–552.

Drew BT, et al. 2014. Another look at the root of the angiosperms reveals a familiar tale. Syst Biol. 63(3):368–382.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5):1792–1797.

Edwards SV, et al. 2016. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol Phylogenet Evol. 94(Pt A):447–462.

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2012. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr Biol. 22(15):1456–1457.

Goremykin VV, et al. 2013. The evolutionary root of flowering plants. Syst Biol. 62(1):50–61.

Graham SW, Iles WJD. 2009. Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. Am J Bot. 96(1):216–227.

Heath TA, Hedtke SM, Hillis DM. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol. 46:239–257.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol Biol Evol. 27(3):570–580.

Jansen RK, et al. 2006. Phylogenetic analyses of *Vitis* (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol. 6:32.

Jansen RK, et al. 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(49):19369–19374.

Judd WS, Campbell CS, Kellogg EA, Stevens PF, Donoghue MJ. 1999. Plant Systematics: a Phylogenetic Approach (Sinauer Associates).

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst Biol. 56(1):17–24.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25(7):971–973.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870.

Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. 34(3):772–773.

Larget BR, Kotha SK, Dewey CN, Ane C. 2010. BUCKy: gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics 26(22):2910–2911.

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(4):611–615.

Laurin-Lemay S, Brinkmann H, Philippe H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr Biol. 22:R593–R594.

Lee EK, et al. 2011. A functional phylogenomic view of the seed plants. PLoS Genet. 7:e1002411.

Leebens-Mack J, et al. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol. 22(10):1948–1963.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24(21):2542–2543.

Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. Syst Biol. 58(4):452–460.

Liu L, Yu L, Pearl DK, Edwards SV. 2009. Estimating species phylogenies using coalescence times among sequences. Syst Biol. 58(5):468–477.

Liu L, Yu L, Edwards SV. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol Biol. 10:302.

Mathews S, Donoghue MJ. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science 286(5441):947–950.

Mirarab S, et al. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30(17):541–548.

Mirarab S, Warnow T. 2015. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52.

Mirarab S. 2017. Phylogenomics: constrained gene tree inference. Nat Ecol Evol. 1(2):0056.

Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A. 104(49):19363–19268.

Moore MJ, et al. 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(4):541–558.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Qiu Y-L, et al. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J Syst Evol. 48(6):391–425.

Roch S, Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst Biol. 64(4):663–676.

Ruhfel BR, Gitzendanner MA, Soltis PS, Soltis DE, Burleigh JG. 2014. From algae to angiosperms – inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. BMC Evol Biol. 14:23.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7):1654–1668.

Seo T-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol Biol Evol. 25(5):960–971.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. 1(5):0126.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17(12):1246–1247.

Shimodaira H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 51(3):492–508.

Simmons MP, Gatesy J. 2015. Coalescence vs. concatenation: sophisticated analyses vs. first principles applied to rooting the angiosperms. Mol Phylogenet Evol. 91:98–122.

Simmons MP. 2017. Mutually exclusive phylogenomic inferences at the root of the angiosperms: *Amborella* is supported as sister and observed variability is biased. Cladistics 33(5):488–512.

Soltis DE, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, rbcL, and atpB sequences. Bot J Linn Soc. 133(4):381–461.

Soltis DE, et al. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am J Bot. 98(4):704–730.

Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402(6760):402–404.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.

Stefanovié S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol Biol. 4:35.

Wang H, et al. 2009. Rosid radiation and the rapid rise of angiosperm-dominated forests. Proc Natl Acad Sci U S A. 106(10):3853–3858.

Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 111:E4859–E4868.

Wiens JJ. 2003. Missing data, incomplete characters and phylogenetic accuracy. Syst Biol. 52(4):528–538.

Xi Z, Liu L, Rest JS, Davis CC. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst Biol. 63(6):919–932.

Yang X, Tuskan GA, Tschaplinski TJ, Cheng Z-M. 2007. Third-codon transversion rate-based *Nymphaea* basal angiosperm phylogeny – concordance with developmental evidence. Nat Preced. 1–20.

Zeng L, et al. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat Commun. 5:4956.

Associate editor: Martin Embley