American Journal of Botany 96(1): 216–227. 2009.

Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny[1]

Sean W. Graham[2] and William J. D. Iles

UBC Botanical Garden & Centre for Plant Research (Faculty of Land and Food Systems), 2357 Main Mall, and Department of Botany, 6270 University Boulevard, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada

We examined multiple plastid genes from a diversity of gymnosperm lineages to explore the consistency of signal among different outgroups for rooting flowering plant phylogeny. For maximum parsimony (MP), most outgroups attach on a branch of the underlying ingroup tree that leads to *Amborella*. Maximum likelihood (ML) analyses either root angiosperms on a nearby branch or find split support for these neighboring root placements, depending on the outgroup. The inclusion of two species of Hydatellaceae, recently recognized as an ancient line of angiosperms, does not aid in inference of the root. Cost profiles for placing the root in suboptimal locations are highly correlated across most outgroup comparisons, even comparing MP and ML profiles. Those for Gnetales are the most deviant of all those considered. This divergent outgroup either attaches on a long eudicot branch with moderate bootstrap support in MP analyses or supports no particular root location in ML analysis. Removing the most rapidly evolving sites in rate classifications based on two divergent angiosperm root placements with Gnetales yields strongly conflicting root placements in MP analysis, despite substantial overlap in the estimated sets of conservative sites. However, the generally high consistency in rooting signal among distantly related gymnosperm clades suggests that the long branch connecting angiosperms to their extant relatives may not interfere substantially with inference of the angiosperm root.

Key words: *Amborella*; angiosperm phylogeny; gymnosperm outgroups; Nymphaeales; rate classes; root cost profiles; taxonomic sampling; tree rooting; *Trithuria*; water lilies.

The radical rerooting of the crown clade of flowering plants around the branch leading to *Amborella trichopoda*, Amborellaceae (and perhaps the water lilies, Nymphaeaceae) was arguably one of the most exciting new findings to emerge since the dawn of plant molecular systematics—although for a new generation of plant biologists it has fast become basic textbook knowledge (e.g., Raven et al., 2005; Judd et al., 2008). This realignment was coupled with the recognition that Austrobaileyales, a small group of woody taxa, is the probable sister group of most other angiosperms (/Mesangiospermae; Cantino et al. 2007). The root node defines the ancestor of all living angiosperms and thus indicates the correct order of the deepest splits in crown angiosperm phylogeny, which is critical for reconstructions of character evolution and for estimating the timing and rate of early diversification events. The "circa-*Amborella*" root placement and our revised understanding of the remainder of the angiosperm backbone continue to trigger re-evaluations of comparative data in diverse fields of plant biology, from paleobotany and morphology to ecophysiology and plant genomics (e.g., Friis et al., 2006; Feild and Arens, 2007; Doyle, 2008; Endress, 2008; Soltis et al., 2008; also see various articles in this issue).

A flurry of publications (Mathews and Donoghue, 1999, 2000; Parkinson et al., 1999; Qiu et al., 1999, 2000; Soltis et al., 1999, 2000; Barkman et al., 2000; Graham and Olmstead, 2000a, b; Graham et al., 2000) followed the initial announcement of the reordered angiosperm tree (at the XVI International Botanical Congress, St. Louis, Missouri, USA, 1999). These findings have been widely accepted because of the overall consistency of results among different studies, which represented a broad array of taxon samplings, gene and genome samplings, and analytical methods. Further support and refinement has come from studies that considered additional loci and approaches (e.g., Qiu et al., 2001, 2005, 2006; Zanis et al., 2002; Borsch et al., 2003; Hilu et al., 2003; Nickerson and Drouin, 2004; Müller et al., 2006) and from combined analysis of molecular and morphological data (e.g., Doyle and Endress, 2000). While there is still a dearth of support from morphology for any particular root of angiosperms (although see Doyle, 2008), this is likely a function of the large morphological distance separating extant angiosperms from related seed plants. Our current understanding of angiosperm morphology appears to be readily reconcilable with the revised root of flowering-plant phylogeny (e.g., Endress and Doyle, 2009, pp. 22–66 in this issue). These molecular systematic findings on the early diversification of the extant angiosperms were capped by the recent discovery, based on plastid, nuclear, and morphological data, that Hydatellaceae, a small and poorly known family of diminutive aquatic plants traditionally thought to be monocots, are actually the sister group of the water lilies (Saarela et al., 2007). This result has added a significant new strand to our understanding of the morphological diversity of the lines that descended from the earliest evolutionary splits in the crown angiosperms (e.g., Friis and Crane, 2007; Rudall et al., 2007, 2008; Friedman, 2008; Remizowa et al., 2008; Sokoloff et al., 2008).

The report of a rooting of angiosperms away from *Amborella* and along the line leading to grasses (Goremykin et al., 2003, 2004, 2005) created a subsidiary stir (e.g., Soltis and Soltis, 2004; Soltis et al., 2004; Martin et al., 2005; Lockhart and

[1] Manuscript received 23 September 2008; revision accepted 30 November 2008.

The authors are grateful to T. S. Feild for providing material of *Trithuria filamentosa* and to S. Magallón and S. Stefanović for helpful comments. This work was supported by a University of British Columbia Graduate Fellowship to W.J.D.I. and by an NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery Grant to S.W.G.

[2] Author for correspondence (e-mail: swgraham@interchange.ubc.ca)

doi:10.3732/ajb.0800320

Penny, 2005). However, it quickly became apparent that this minority report was likely a consequence of the small and taxonomically skewed sample of whole plastid genomes available, which evidently caused artefactual attraction between long, undivided branches in the ingroup and outgroup. The branch connecting angiosperms to other seed plants is particularly lengthy, and only one seed plant outgroup (*Pinus*) was available at the time. In addition, one of the most heavily sampled angiosperm clades in these studies, the grasses, is well known to have an elevated substitution rate (e.g., Gaut et al., 1992), resulting in substantially longer branches than for most other groups of monocots (e.g., Graham et al., 2006). All subsequent phylogenetic analyses of angiosperms using whole plastid genomes have had denser samplings, including improved samplings of monocots and other key lineages absent in the studies of Goremykin et al. (2003, 2004, 2005). They consistently uphold a root on the branch leading to *Amborella*, or to *Amborella* and water lilies, using a variety of analytical methods (e.g., Stefanović et al., 2004; Leebens-Mack et al., 2005; Cai et al., 2006; Hansen et al., 2007; Jansen et al., 2007; Moore et al., 2007; Mardanov et al., 2008). This supports the conjecture of Soltis and Soltis (2004), Stefanović et al. (2004) and Leebens-Mack et al. (2005) that the grass rooting of angiosperms seen by Goremykin and colleagues was primarily a function of strong long-branch attraction between ingroup and outgroup taxa (see Felsenstein, 1978; Hendy and Penny, 1989), in turn a consequence of too-limited taxonomic sampling.

The high degree of congruence concerning the root of flowering plant phylogeny from nearly all molecular phylogenetic studies is reassuring. Nonetheless, we should pay attention to the possibility that the broad convergence in analyses of available data might mask subtle (or perhaps not so subtle) pathological behavior in inferences of the root of angiosperm phylogeny, reflecting imperfectly understood long branch effects, or imperfectly modeled molecular evolution on these branches (e.g., Matsen and Steel, 2007). In simulation studies of rapid radiations subtended by long outgroup branches, Holland et al. (2003) and Shavit et al. (2007) pointed out that arbitrarily long DNA sequence alignments can give misleading results, even when the methods used are statistically consistent (note that consistency is a property of infinitely large data sets). They also reminded us that the mathematical behavior of long branches is still poorly understood outside relatively limited analytical examples involving a handful of taxa. In general, rapid radiations subtended by long outgroup branches can be difficult to resolve satisfactorily (e.g., Graham et al., 2002; Rodríguez-Ezpeleta et al., 2007; Murdock, 2008). For angiosperms, Zanis et al. (2002) noted the possibility that some outgroups may have a higher ratio of noise to phylogenetic signal than others, and at least one of the major (and the longest of all) outgroup branches, Gnetales, is well known to "misbehave" in seed plant phylogenetic inference using molecular data (e.g., Burleigh and Mathews, 2004; Mathews, 2009, pp. 228–236 in this issue). This raises the possibility that this outgroup (at least) might provide aberrant estimates of the root of angiosperm phylogeny. Even slight systematic bias may become magnified in these situations (Hedtke et al., 2006).

Simulation studies have generally shown that dense taxon sampling helps to minimize long-branch attraction artefacts and improves accuracy (e.g., Hillis, 1998; Zwickl and Hillis, 2002; Hillis et al., 2003; Hedtke et al., 2006). However, there are well-defined upper limits as to how far taxonomic sampling can be improved in molecular studies regarding the root of flowering plant phylogeny, because extinction has significantly pruned much of the early diversity of angiosperms and of the major seed plant clades in general (e.g., Rothwell and Stockey, 2002; Crane et al., 2004; Mathews, 2009). Arguably, most molecular studies that have addressed the issue do have an appropriate sampling of the diversity of the early crown angiosperm lines. However, it is surprising that no study to date has systematically examined the effect of variable outgroup sampling. This should be a point of some concern because most published studies have a relatively low density of the few gymnosperm lines that have persisted to the modern day (e.g., currently no more than three for the whole plastid genome studies cited earlier). While a low number of outgroups (perhaps even one) may prove to be adequate, the consistency of the signal among available outgroups regarding the placement of the angiosperm root would benefit from further study.

Since we first published on the question of the root of the flowering plants (Graham and Olmstead, 2000b; Graham et al., 2000), we have accumulated a large number of outgroup taxa for a comparable gene sampling (17 plastid genes), representing all of the major lineages (cycads, *Ginkgo*, Gnetales, and two major lines of conifers, Pinaceae and /Cupressophyta; the slash before the latter major clade of conifers indicates that it is a non-Linnean name; Cantino et al., 2007) with reasonably extensive sampling densities in most cases (Rai et al., 2003, 2008; Zgurski et al., 2008). In principle, different gymnosperm outgroups provide at least partly independent estimates of the root of flowering plant phylogeny, because the path (sum of branches) between a given outgroup terminal and the angiosperm root is only partly shared with other outgroups (see Graham et al., 2002, for a comparable example in monocots). We therefore use different outgroup combinations here to assess whether they provide substantially different signals concerning the root of angiosperm phylogeny and whether adding a dense sampling improves our confidence in the results. We also included an additional member of Hydatellaceae and examined the effect of removing rapidly evolving characters in a likelihood-based classification of site rates for one outgroup that is potentially especially problematic, Gnetales.

MATERIALS AND METHODS

Taxonomic and genomic sampling—The plastid matrix considered here includes 66 seed plant representatives (30 angiosperms and 36 gymnosperms), representing all major lineages. The regions examined are those considered in Saarela et al. (2007), comprising *atpB*, *rbcL*, 10 photosystem II genes (*psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbJ*, *psbL*, *psbN*, *psbT*), three ribosomal protein genes (*rpl2*, 3′-*rps12*, *rpl2*), two NADH dehydrogenase subunit genes (*ndhB* and *ndhF*), and several conservative noncoding regions (one intron each in *ndhB*, *rpl2*, and 3′-*rps12*; the intergenic spacers between 3′-*rps12* and *rps7*, and between *ndhB* and *trnL*[CAA]). Collectively, these represent approximately one tenth of the plastid genome. The final matrix is included as online material (Appendix S1, see Supplemental Data with the online version of this article). Sequences from *Trithuria filamentosa* Rodway, Hydatellaceae (*T. Feild 210* (TENN); GenBank numbers (FJ514801–FJ514808) are new here; source details and GenBank numbers for the other taxa are in Graham and Olmstead (2000a, b); Graham et al. (2006), Rai et al. (2003, 2008), Saarela et al. (2007) and Zgurski et al. (2008). Methods of DNA extraction, amplification, sequencing, and alignment follow Graham and Olmstead (2000b), Graham et al. (2000, 2006) and Rai et al. (2003, 2008).

Phylogenetic analyses—We used PAUP* version 4.0b10 (Swofford, 2002) for the maximum parsimony (MP) analyses, performing heuristic MP tree searches with 100 random addition replicates, but otherwise using the default settings in the program PAUP* (e.g., tree-bisection-reconnection [TBR] branch-swapping). For ML analysis, we determined the optimal maximum likelihood (ML) model for the angiosperms using the hierarchical likelihood ratio test (hLRT) and the Akaike information criterion (AIC), as implemented in the program ModelTest version 3.7 (Posada and Crandall, 1998). The GTR + Γ + I model was supported as the optimal DNA substitution model by both methods (this model assumes a general-time-reversible [GTR] rate matrix with the proportion of invariable sites [I] considered, and with among-site rate variation accounted for using the gamma [Γ] distribution). Because there is unlikely to be a decrease in model complexity with the addition of distant outgroup taxa, we used this model for all ML analyses performed here, estimating model parameters during ML analysis, but using empirical estimates of the base frequencies. We used the program PhyML version 3.0 (an updated version of 2.4.4, Guindon and Gascuel, 2003, website at http://www.atgc-montpellier.fr/phyml) for ML heuristic tree searching with default search settings (i.e., a BIONJ starting tree and nearest-neighbor interchange [NNI] branch-swapping) and used PAUP* for the ML versions of the tree rooting experiments outlined later.

Bootstrap support for optimal and nearly optimal root placements—We estimated MP and ML branch support using nonparametric bootstrap analysis (Felsenstein, 1985) for 100 bootstrap replicates, using the general search settings noted already, but with a single random addition replicate per bootstrap replicate for parsimony analysis. For branches of interest with less than 50% bootstrap support, we inferred support values from the bootstrap bipartition log in PAUP*. We generated a comparable log for the ML analyses by converting the PhyML bootstrap tree output to Newick format to allow calculation of a majority-rule consensus in PAUP*. We repeated tree searches using a variety of outgroup combinations. We employed as outgroup taxa, respectively, 10 cycads, 18 /Cupressophyta conifers (i.e., all sampled conifers excluding Pinaceae), three Gnetales, *Ginkgo* alone, four Pinaceae, and all 36 gymnosperms together; all analyses included the same 30 ingroup (angiosperm) taxa. We used these different combinations to determine whether the choice of outgroup had an effect on the inference of the optimal root, and with what support.

Shimodaira–Hasegawa tests of optimal and nearly optimal root placements—The underlying angiosperm topology inferred with MP and ML with all 36 outgroups included is identical, when pruned of these outgroups, to the MP tree inferred with no outgroups (see Results). We rerooted this angiosperm topology to examine whether several nearly optimal roots were significantly worse than the optimal one using the Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa, 1999) implemented in PAUP* to compare the resulting tree scores (see Saarela et al., 2007, for details). We repeated this comparison for six different outgroups (cycads, /Cupressophyta, Gnetales, *Ginkgo*, Pinaceae, all gymnosperms). The outgroup subtree topologies ("clades") assumed in each case are as follows: the cycad subtree in Zgurski et al. (2008, their Fig. 2) pruned of two floating taxa (*Bowenia* and *Stangeria*); the /Cupressophyta subtree in Rai et al. (2008), *Ginkgo* by itself, the Gnetales and Pinaceae subtrees in Rai et al. (2008), and the entire gymnosperm subtree shown in Fig. 2 here.

Generating cost profiles for all possible roots of the 30-taxon subtree—We also assessed the cost of forcing the angiosperm root to increasingly suboptimal locations for all possible branches on the unrooted subtree noted before, repeating these analyses using two different optimality criteria (MP vs. ML) for the six different outgroups, and for two MP comparisons using Gnetales where we focused on conservatively evolving sites (discussed later). We first generated the 57 possible angiosperm roots for the 30-taxon unrooted angiosperm tree [number of possible roots = (2 × number of taxa) – 3 branches] using the "All rootings" option in MacClade v. 4.03 (Maddison and Maddison 2001). Each tree file (containing the 57 possible roots) was then edited so that the spliced stem branch of an outgroup subtree was reattached to the different root nodes, with tree files set up for each of the six outgroups considered (i.e., the same outgroup subtrees used for the SH tests). Trees generated from MacClade are truly rooted (i.e., time-irreversible), and so we converted them to "unrooted" trees in PAUP* to permit the use of time-reversible criteria such as the GTR DNA substitution model for ML analysis and unordered (Fitch) parsimony. For these criteria, the root of a given subtree is a matter of perspective, given external knowledge about how the presumptive ingroup and outgroups attach to the rest of the tree of life (PAUP* and PhyML make no assumptions about where the root of a given taxon set lies during tree-searching using time-reversible criteria, even when an outgroup is "defined."). However, since we can safely assume that the root of the seed-plant tree as a whole does not lie within extant angiosperms, the point of connection between angiosperms and outgroups (gymnosperms) provides certainty over the local arrow of time, and so it defines the root of the angiosperms as a whole.

Effect of removing rapid sites on root inference using Gnetales—As Gnetales are a particularly divergent seed plant outgroup (see Results), we explored whether rooting preferences using this three-taxon outgroup were substantially affected by removing sites inferred to be very rapidly evolving, at least according to a ML-based rate method for classifying the rate class of each aligned site. We employed the program HyPhy version 0.99 beta (Kosakovsky Pond et al., 2005; version released on 8 December 2006) to partition the matrix into different rate classes, given a GTR model and using two different user-supplied trees. As two extremes, we considered site-rate classifications based on the angiosperm subtree considered earlier, but rooted with Gnetales at the optimal MP root of angiosperms in this case (i.e., a long terminal branch leading to *Arabidopsis*; see Results), or at the branch leading to *Amborella*. HyPhy assigns each aligned nucleotide to its most likely individual rate category (e.g., for nine classes, RC0 represents sites with no change and RC8 the fastest sites). We deleted the two fastest rate classes (RC7 and RC8) and re-ran the MP analyses described using the remaining (and more conservative) characters (i.e., RC0–RC6).

Generating and comparing root cost profiles—Each tree file (representing a given outgroup choice) was input in PAUP* in turn. We used the "Tree scores" option to calculate the cost (tree length) for all 57 possible roots in that profile, repeating this for MP and ML. Logs of root scores were edited and input into the program Excel (Microsoft, Redmond, Washington, USA). We visually compared a given pair of profiles of these root cost profiles (sets of 57 MP or ML tree scores, representing multiple points of attachment of a particular outgroup to the ingroup) using bivariate scatter plots. For example, we compared the profile of ML rooting costs along different parts of the angiosperm backbone using Gnetales as an outgroup vs. using *Ginkgo* as an outgroup. We also calculated the correlation coefficient, *r*, for each pair of root cost profiles (see Graham et al., 2002). For six outgroups and two phylogenetic criteria (MP and ML) plus the two MP comparisons for Gnetales made after removing rapidly evolving sites, there are (14 × 14) – 14 = 182 unique pairwise comparisons of root cost profiles. We depict a subset of these as scatter plots below. This subset was chosen to display the range of correlation scores that we observed. Using (1 – *r*) as a pairwise distance metric between root cost profiles, we then inferred a neighbor-joining (NJ) tree using the Neighbor module of the program PHYLIP 3.67 (Felsenstein, 2007) to visually summarize the overall relationship among all 182 pairwise comparisons.

RESULTS

When the trees recovered from MP analyses using different outgroups are pruned of all outgroups, they yield a single angiosperm subtree topology (a rooted version of this tree is shown in Fig. 1). This tree places monocots as the sister group of (eudicots + *Ceratophyllum*), and Chloranthaceae as the sister group of magnoliids. Chloranthaceae plus magnoliids were then sister to the clade consisting of monocots, eudicots, and *Ceratophyllum*. Almost all the branches in this tree have consistently moderate (≥70%) to strong (≥90%) bootstrap support from MP analysis across all outgroup combinations (most not presented here, but values for the case with all 36 outgroups included are shown in Fig. 1, along with the results when no outgroups were included). Most of the ML analyses using different outgroup combinations inferred an angiosperm subtree that was identical to this tree, except that it depicts (eudicots + *Ceratophyllum*) as the sister group of (magnoliids + Chloranthaceae), with monocots sister to all of these. However, these arrangements were poorly supported by all ML bootstrap analyses (not shown), and the ML analysis that include all 36 gymnosperm outgroups recovered the core angiosperm topology seen in the MP analyses (Fig. 2), albeit rooted in a different position from the equivalent MP analysis using these outgroups (Fig. 1).

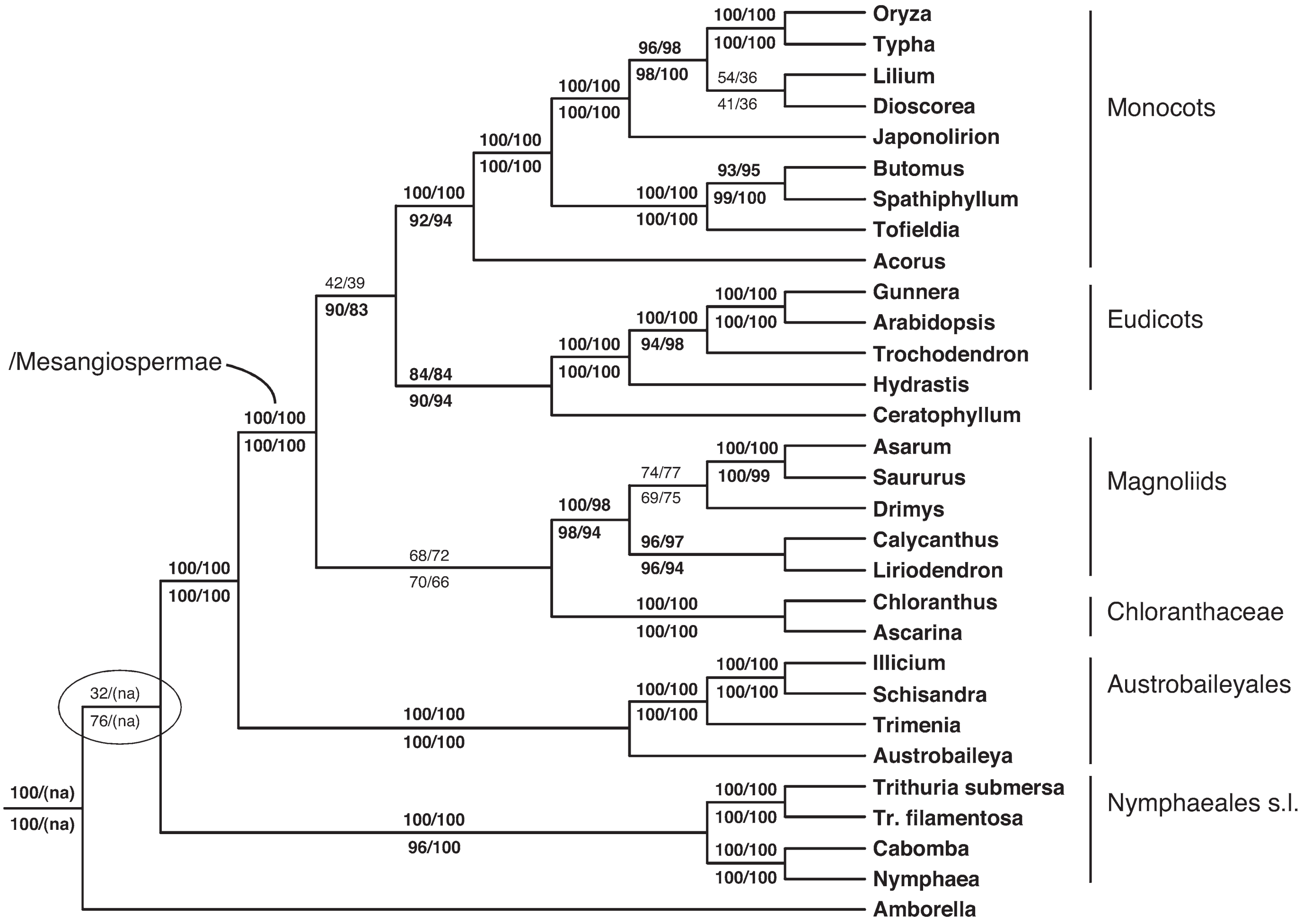


Fig. 1. Rooted maximum parsimony (MP) phylogeny of flowering plants inferred from 17 protein-coding plastid genes and six associated noncoding regions using all 36 outgroups depicted in Fig. 2 (all trimmed here). Bootstrap values above branches from maximum likelihood (ML) analysis; those below branches from MP analysis (left: with outgroups included; right: outgroups excluded). Strongly supported branches (≥90% support) indicated in bolder font. Two branches around root are absent in unrooted tree (na: bootstrap support not applicable). The optimal rooted ML tree for angiosperms is identical to this topology, apart from placement of root (note circled values). For comparison, the ML root arrangement (depicted in Fig. 2) has 55% ML bootstrap support and negligible MP bootstrap support (see Table 1).

Bootstrap support for major outgroup branches in the analysis that included all 66 taxa was generally very strong (for ML bootstrap values of major clades see Fig. 2). However, the position of Gnetales relative to the two major clades of conifers, Pinaceae and /Cupressophyta, was only weakly supported. The best ML tree placed Gnetales as the sister group of /Cupressophyta, with <50% support (Fig. 2). In contrast, the best MP tree (not shown in full) placed Gnetales as the sister group of Pinaceae, with moderate support (71% for the branch connecting Pinaceae and Gnetales, with the monophyly of both taxa well supported). The branches in Gnetales are among the longest in the seed plants, rivaled in length only by the well-supported branch connecting angiosperms to gymnosperms (which is marked as the "angiosperm stem branch" here; Fig. 2). Most branches within the angiosperms were also well supported (boldface numbers in Fig. 1).

Levels of bootstrap support in angiosperms were generally highly consistent across analyses, regardless of the phylogenetic criterion (MP vs. ML; Fig. 1), and regardless of whether outgroups were included or not (e.g., Fig. 1). Apart from the relative arrangement of eudicots, *Ceratophyllum* and monocots (discussed before), the main inconsistency concerns the position of the root of the angiosperms. In most parsimony analyses, *Amborella* was inferred to be the sister group of the remaining angiosperms (see underlined values in Table 1), generally with moderate to strong support (Table 1, Fig. 1), with the next best-supported root placing Nymphaeales s.l. (water lilies + Hydatellaceae) as the sister group of the remaining angiosperms. In contrast, most ML analyses yielded an optimal tree in which a clade consisting of *Amborella* + Nymphaeales s.l. was the sister group of the remaining angiosperms. The distinction between the bootstrap support values for the optimal and next-best root placements for the ML analysis was generally less sharp than was the case for MP analysis (Table 1).

Shimodaira–Hasegawa tests comparing nearly optimal roots—Although the (*Amborella* + Nymphaeales s.l.) root was typically the best of the three for ML, it was statistically indistinguishable from the next two best roots according to the SH test, that is, when compared either to a clade consisting of

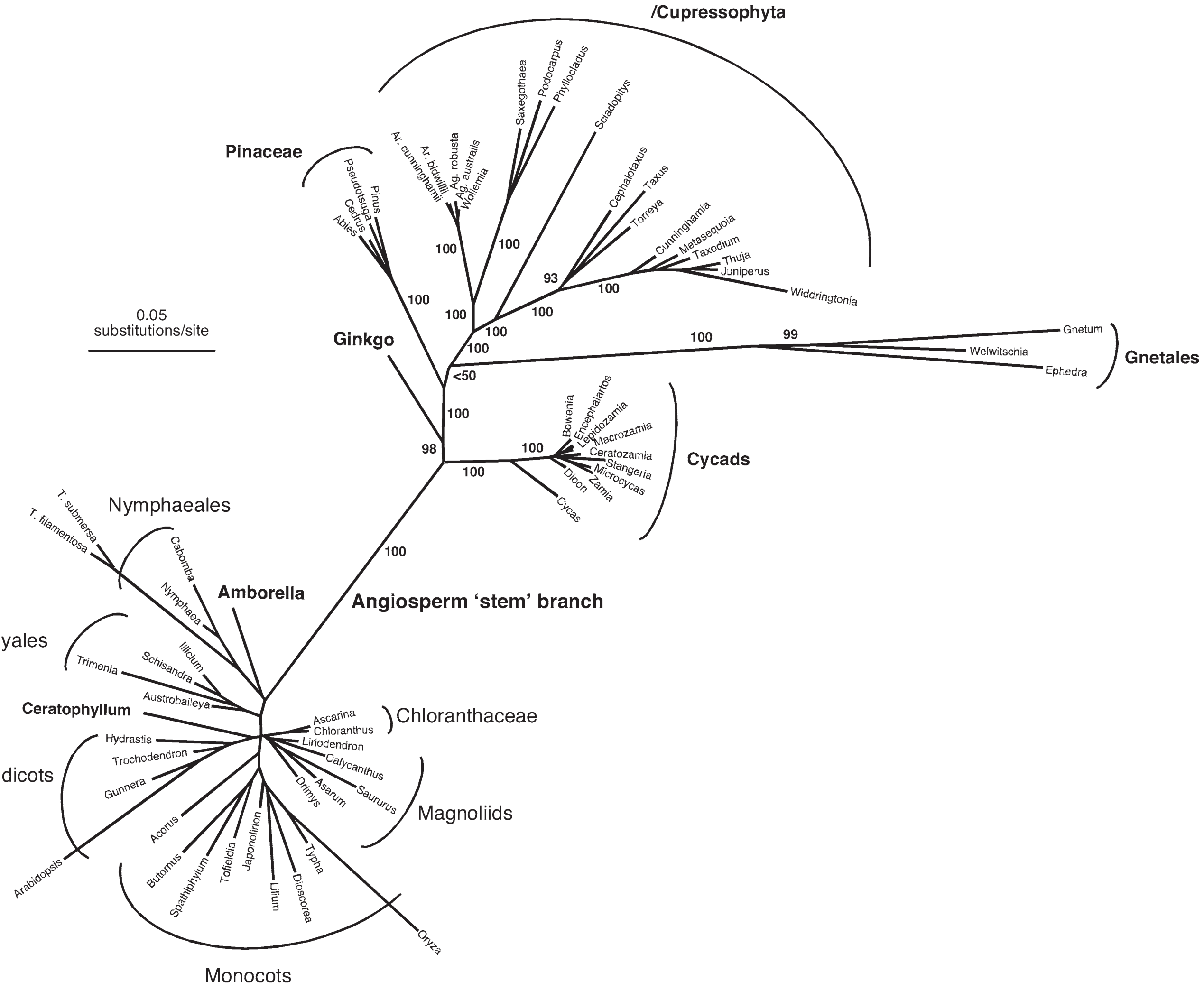


Fig. 2. Unrooted maximum likelihood (ML) phylogeny of flowering plants (angiosperms) and gymnosperms, inferred from 17 protein-coding plastid genes and six associated noncoding regions (−ln L = 143 436.87 [PAUP*] or 143 689.39 [PhyML]; branch lengths computed using PhyML). ML bootstrap support values for major branches indicated beside branches (see Fig. 1 for values in angiosperms). Corresponding maximum parsimony (MP) trees (not shown) are identical to this tree, apart from minor details in cycads and Podocarpaceae, the position of Gnetales (Gnetales are sister to Pinaceae for MP), and the root of angiosperms (*Amborella* sister to remaining angiosperms for MP; Fig. 1). Angiosperm stem branch ("stem" assumes that root of extant seed plants is located elsewhere among gymnosperms) and major clades of angiosperms are marked.

Nymphaeales s.l. as sister to the remaining angiosperms or to *Amborella* alone (Table 2). Other suboptimal root placements were statistically worse, but most outgroup combinations had trouble distinguishing Austrobaileyales as a significantly worse place to root the angiosperms than the optimal placement (Table 2). This root was only definitively rejected when all 33 outgroups were included as outgroups (Table 2). The relative penalties for the suboptimal root positions considered in the SH tests were generally smaller using Gnetales as an outgroup than for any other combination considered here (Table 2).

Comparing root cost profiles across the ingroup tree—Considering all 57 forced root placements on the core angiosperm subtree, there was a very strong linear relationship between the costs inferred using different outgroups for a given optimality criterion (e.g., Figs. 3, 5, 6; these specific comparisons were chosen to visually demonstrate the range among the many pairwise comparisons made). All pairwise correlations in root cost profiles were significant, but the weakest correlations were observed when one of the root profiles under consideration was for Gnetales (e.g., Figs. 7–10; note how these plots show considerably more scatter than Figs. 3–6). These include significant correlations in comparisons of each outgroup but using different optimality criteria (MP vs. ML, e.g., Figs. 4 and 8 for *Ginkgo* and Gnetales, respectively), although the amount of scatter in the MP vs. ML comparison was again greatest for Gnetales, of all outgroups considered here (Fig. 8 and see the pairwise MP : ML distances for each outgroup taxon in Fig. 11). The two outliers in all of the bivariate scatter plots represent forced root placements on each of two sampled branches in Hydatellaceae. These branches apparently define a relatively recent split, with the consequence that a high penalty must be incurred to place the angiosperm root inside the family. Removing these two outlier branches has essentially no effect on any of the pairwise correlations estimated here (data not shown).

Table 1. Bootstrap support for three optimal or nearly optimal roots of flowering plants according to various outgroups, using maximum likelihood (ML) and parsimony (MP); optimal root in corresponding tree-rooting experiments is underlined. Unless noted, there is 100% bootstrap support for monophyly of angiosperms, for Nym s.l. (Nymphaeales s.l. = water lilies and Hydatellaceae), and for clade comprising /Mesangiospermae and Austrobaileyales. Amb = *Amborella*; "–" = Not seen in bootstrap log (<5%). See text for outgroup composition in each case.

	Sister lineage to the remaining angiosperms		
Outgroup used (phylogenetic criterion)	(Amb+Nym s.l.)	Amb	Nym s.l.
All outgroups (ML)	55%	32%	13%
All outgroups (MP)	—	76%	≤24% [a]
Cycads (ML)	32%	57%	11%
Cycads (MP)	—	81%	≤19% [a]
/Cupressophyta (ML)	80%	20%	—
/Cupressophyta (MP)	—	93%	5% [a]
Gnetales (ML) [b]	≤8% [c]	7%	—
Gnetales (MP) [b]	—	—	—
Ginkgo (ML)	45%	38%	17%
Ginkgo (MP)	—	84%	≤16% [a]
Pinaceae (ML)	85%	13%	2%
Pinaceae (MP)	—	66%	≤32% [a]

[a] Maximum values in these cases, because monophyly of Nymphaeales s.l. (= water lilies and Hydatellaceae) has <100% MP bootstrap support here (range: 70–96%); support for clade consisting of /Mesangiospermae and Austrobaileyales ranges from 98 to 100% in these cases.

[b] Best supported root unclear for ML here, because of weak support for multiple clades around root of angiosperms. For MP, best-supported root places *Arabidopsis* as sister to other angiosperms, with 70% bootstrap support.

[c] Maximum value.

The NJ tree shown in Fig. 11 summarizes pairwise comparisons among all 182 possible comparisons. Most comparisons fall close to the range shown in Figs. 3–6, except for those involving Gnetales (e.g., note correlation coefficients on individual plots and compare to correlation-based distances in Fig. 11). ML profiles involving all of the outgroups behave virtually in lock-step in terms of rooting preferences, except for those based on Gnetales (Fig. 11, and compare the example scatter-plots shown in Figs. 5 or 6 to Figs. 9 or 10, respectively). The pairwise comparisons among the MP root profiles involving most outgroup combinations were also very similar, but Gnetales were again a substantial outlier (Fig. 11, and compare Figs. 3 and 7). However, a given ML cost profile for Gnetales was less weakly correlated to those of other outgroups (for ML), than the same MP root profile was to corresponding cases for MP (Fig. 11).

The best MP root found using Gnetales as the outgroup was on the long branch leading to *Arabidopsis* (i.e., within the eudicots; not shown). This unusual arrangement was moderately well supported by bootstrap analysis (70% support). In contrast, the best ML root for this outgroup was on the branch leading to *Amborella* (one of three marked in Figs. 9, 10), although there was very poor ML bootstrap support for any particular relationship among the major clades of angiosperms in this case (Table 1).

Rate classifications and root inference using Gnetales—When the fastest site-rate classes (RC78) inferred using Gnetales as an outgroup were removed from consideration, the resulting MP root profiles (i.e., for RC06 based on either the *Amborella* root favored in ML analysis using Gnetales or the improbable *Arabidopsis* root favored in MP analysis) were substantially more similar to those inferred with ML analysis (for any outgroup), than to the MP analysis with all characters included for Gnetales (Fig. 11). Despite this apparent convergence, the two different RC06 rate classifications yielded strongly supported but conflicting MP trees (not shown). The RC06 classification based on the Gnetales root at *Amborella* strongly favored an *Amborella* root placement for this outgroup (95% MP bootstrap support), whereas one based on the *Arabidopsis* rooting strongly favored an *Arabidopsis* root (98% MP support). It appears that relatively few characters led to inference of these contrasting arrangements; of 1406 variable characters found across the two RC06 classifications, only 13 and 20 sites are unique to each one (i.e., for classifications based on the a priori Gnetales root assignments for *Amborella* vs. *Arabidopsis* roots, respectively).

DISCUSSION

The especially long branch separating extant angiosperms and gymnosperms in all molecular phylogenies is a function of the extended time since their most recent common ancestor, the repeated thinning of intervening lineages by extinction and perhaps also of elevated substitution rates along this branch. Graham et al. (2002) noted that increasing the sampling among the most closely related outgroups should increase our ability to discriminate the root of an ingroup, when outgroups and ingroup are distantly related to each other (see also Shavit et al., 2007); this is not unexpected given the general improvement expected in phylogenetic accuracy with denser taxon sampling (e.g., Zwickl and Hillis, 2002). However, given the small number of major gymnosperm lines that have survived to the present day, there may be limits to the improvement that is possible here (e.g., Mathews, 2009). Nonetheless, there are still four or five major seed plant outgroups that can be used as somewhat independent estimators of the angiosperm root. They can be considered partly independent because a good fraction of the homoplasy that might mislead rooting the ingroup tree should have accumulated in different ways for each outgroup, on the divergent branches neighboring the long angiosperm stem branch (marked in Fig. 2). Homoplasy decreases phylogenetic information (by removing states at sites that would have been informative about the correct relationship) and also potentially increases misinformation (due to the accumulation of parallel substitutions). The latter effect is more problematic because it may lead to long-branch attraction, but we might expect it to affect different outgroups in somewhat different ways, assuming independent substitutions following divergence from common outgroup ancestors. A first step in determining whether there may be bias in the inference of the angiosperm root is therefore to characterize rooting preferences for different outgroups. Encouragingly, the comparisons performed here show that there is significant correlation among all the root cost profiles, regardless of outgroup or phylogenetic criterion (MP vs. ML).

More remarkably, for most of the outgroup comparisons the root cost profiles are nearly perfectly correlated when we use ML as the optimality criterion (Figs. 5, 6, 11). We would expect ML to be less affected by long-branch problems than MP (e.g., Anderson and Swofford, 2004). However, in most cases there is also barely any more scatter in rooting preferences between outgroups when using MP alone (see Fig. 3 for an example).

Table 2. Shimodaira–Hasegawa tests of whether suboptimal roots of flowering plants are significantly different from optimal root one, according to maximum-likelihood comparisons that consider various seed plant outgroups. Amb = *Amborella*; Nym s.l. = Nymphaeales s.l. (= water lilies + Hydatellaceae); Hyd = Hydatellaceae; Nym s.s. = water lilies; Aust = Austrobaileyales. See text for outgroup composition in each case.

	Increase in −ln L for alternative root placements					
Outgroup (−ln L of best)	(Amb+Nym s.l.)	Amb	Nym s.l.	Hyd	Nym s.s.	Aust
All outgroups (143436.87)	**Best root**	2.00	3.32	95.28**	95.29**	45.09*
Cycads (85752.39)	**Best root**	0.76	2.40	77.95**	77.95**	30.93^
/Cupressophyta (115549.03)	**Best root**	2.63	6.04	85.01**	85.23**	27.25^
Gnetales (83143.54)	0.601	**Best root**	1.43	39.12**	39.12**	14.05
Ginkgo (79429.04)	**Best root**	1.16	1.62	65.10**	65.10**	24.14^
Pinaceae (82257.26)	**Best root**	3.21	3.36	59.03**	59.03**	16.38

Notes: ** $P < 0.001$; * $P < 0.05$. ^: $P \leq 0.10$

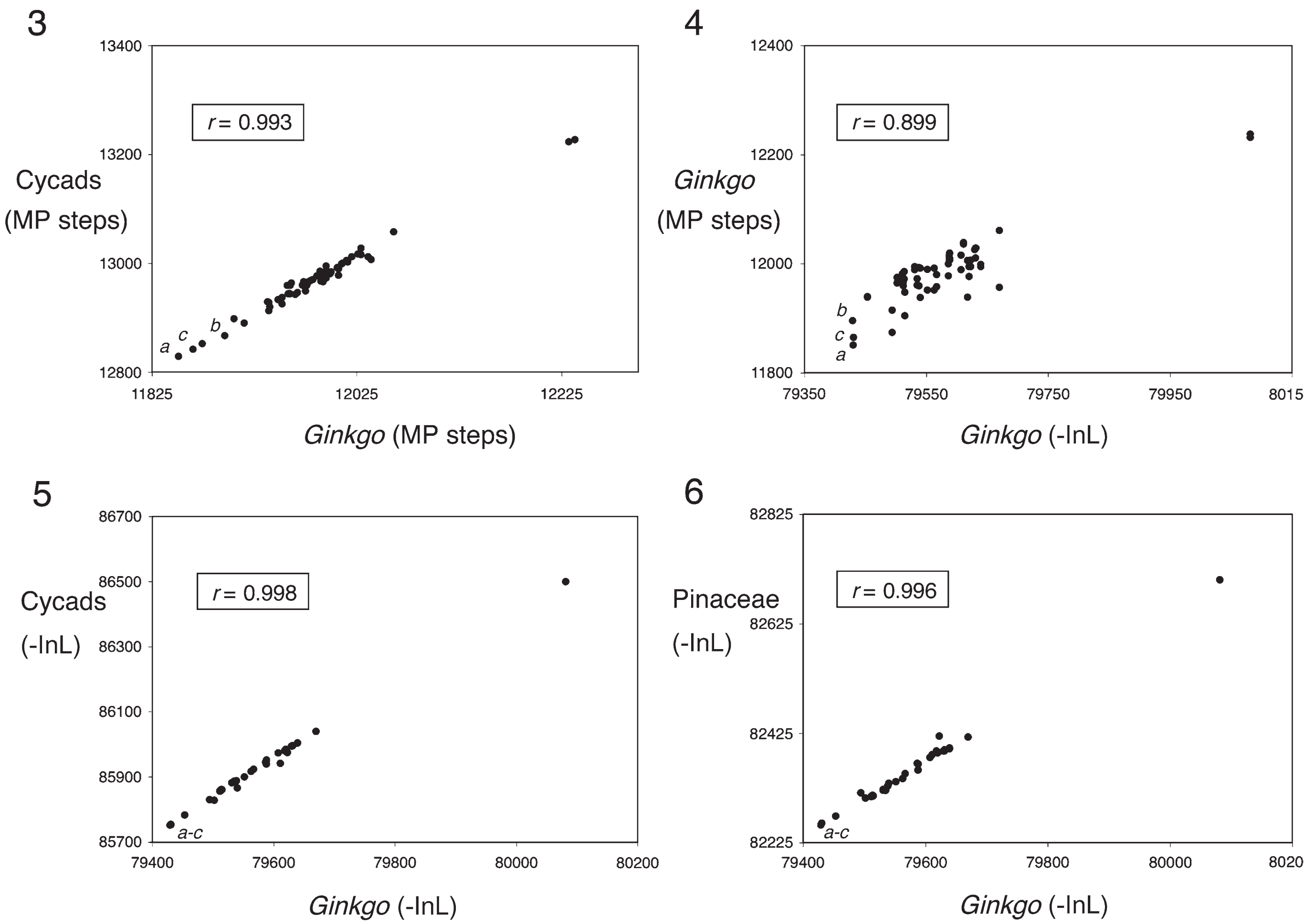
The modest divergence in rooting preferences between MP and ML for a given outgroup (Fig. 11, and see Fig. 4 for an example) indicates that parsimony and likelihood "see" the angiosperm root in somewhat different ways, even though they largely agree on what they "prefer." This is also reflected in distinctive behavior of these two optimality criteria in terms of bootstrap support for the angiosperm root. Ignoring Gnetales for the moment, the MP bootstrap analyses tend to favor a single root, the *Amborella* root, with moderate to strong support (Table 1). In contrast, the ML bootstrap analyses are frequently more "agnostic," with little to separate an *Amborella* root or a root on the lineage leading to (*Amborella* and Nymphaeales s.l.), at least when cycads, *Ginkgo*, or the combination involving all 36 outgroups are used (Table 1). Graham et al. (2002) suggested that outgroups with randomized signal may prefer "terminal" (unbroken ingroup) branches rather than the "internal" branches on an unrooted tree, even when the latter are quite long, which is perhaps an expression of a tendency toward certain tree topologies when outgroups are used to root rapid radiations (Shavit et al., 2007). We wonder if a preference for an *Amborella* root in most MP analyses here might therefore reflect a somewhat greater susceptibility to tree misinference for this optimality criterion. In this regard, it is notable that MP analysis provides no appreciable bootstrap support for the root placement in which a clade consisting of *Amborella* + Nymphaeales s.l. is the sister group of all other angiosperms (Table 1), the root preferred in most ML analyses.

The outgroup that appears to be most divergent in its rooting preferences, Gnetales, has no firm preference in ML analysis (Table 2), but clearly prefers an incorrect root in MP analysis (70% support for an angiosperm root on the branch leading to *Arabidopsis*, one of the longest terminal branches in the ingroup tree here). The "agnosticism" of ML analysis using a Gnetales root seems desirable in this situation.

Removing the most rapidly evolving characters according to likelihood-based site rate classifications improved the correlation of MP root cost profiles for Gnetales as compared to other outgroups (Fig. 11). Unfortunately, however, the starting tree used to estimate these rate classes can have a strongly misleading effect on phylogenetic inference. Each set of rate classes resulted in strong bootstrap support for the root placement that was used to generate the rate classification (i.e., *Arabidopsis* vs. *Amborella*, with 98% and 95% bootstrap support, respectively). Both of these roots cannot be correct, and so at least one of them must be strongly misleading. These contrasting rate classifications inferred nearly the same set of slowly evolving characters (RC06; see Results), but evidently the few characters that they did not share were sufficient to mislead phylogenetic inference using parsimony. It should be of concern that a small handful of characters from a large DNA sequence alignment seems to be responsible for strongly misleading tree inference in this situation. Because we do not know what the correct root is in advance and different model trees affect the ML-based site-rate classifications sufficiently to yield conflicting tree inferences, using these classifications may not be a solution to the misinference of the angiosperm root using Gnetales. Even if it turned out that Gnetales are the sister group of angiosperms among extant plants, a relationship recovered in most morphology-based studies (e.g., Rothwell et al., 2009, pp. 296–322 in this issue, but see Doyle, 2006), we do not recommend using them to root the flowering plant tree using molecular data because of their evident tendency to misinfer the root node of angiosperm phylogeny.

Most outgroups behaved very consistently in the inference of the root node. However, we cannot definitively rule out the possibility that undetected DNA substitutional events on the angiosperm stem branch might have misled our ML analyses. If so, ML models that take account of this complexity could provide a different answer to what we found here. For example, partitioned likelihood analyses may be an improvement over the ML analyses we performed. However, we generally find that different character subpartitions of the plastid regions used here yield comparable phylogenetic results, even for older crown clades (e.g., cycads; Zgurski et al., 2008), and so we did not extensively explore this possibility. Nonetheless, we re-examined our correlation results for one outgroup combination (all 36 outgroup taxa considered simultaneously) by considering whether the first two codon positions behave substantially differently than the third codon position. These two data partitions exemplify quite distinct DNA substitutional dynamics. The former data partition tends to reflect nonsynonymous substitutions and is often considered to be less "saturated" than the latter data partition, which largely comprises synonymous substitutions (e.g., Sanderson et al., 2000). Despite these differences, a circa-*Amborella* root was the optimal one for both data partitions; they marginally preferred a root placing *Amborella* and Nymphaeales as sister to other angiosperms (data not shown). The two root cost profiles based on these codon partitions were also highly correlated to each other ($r = 0.853$).

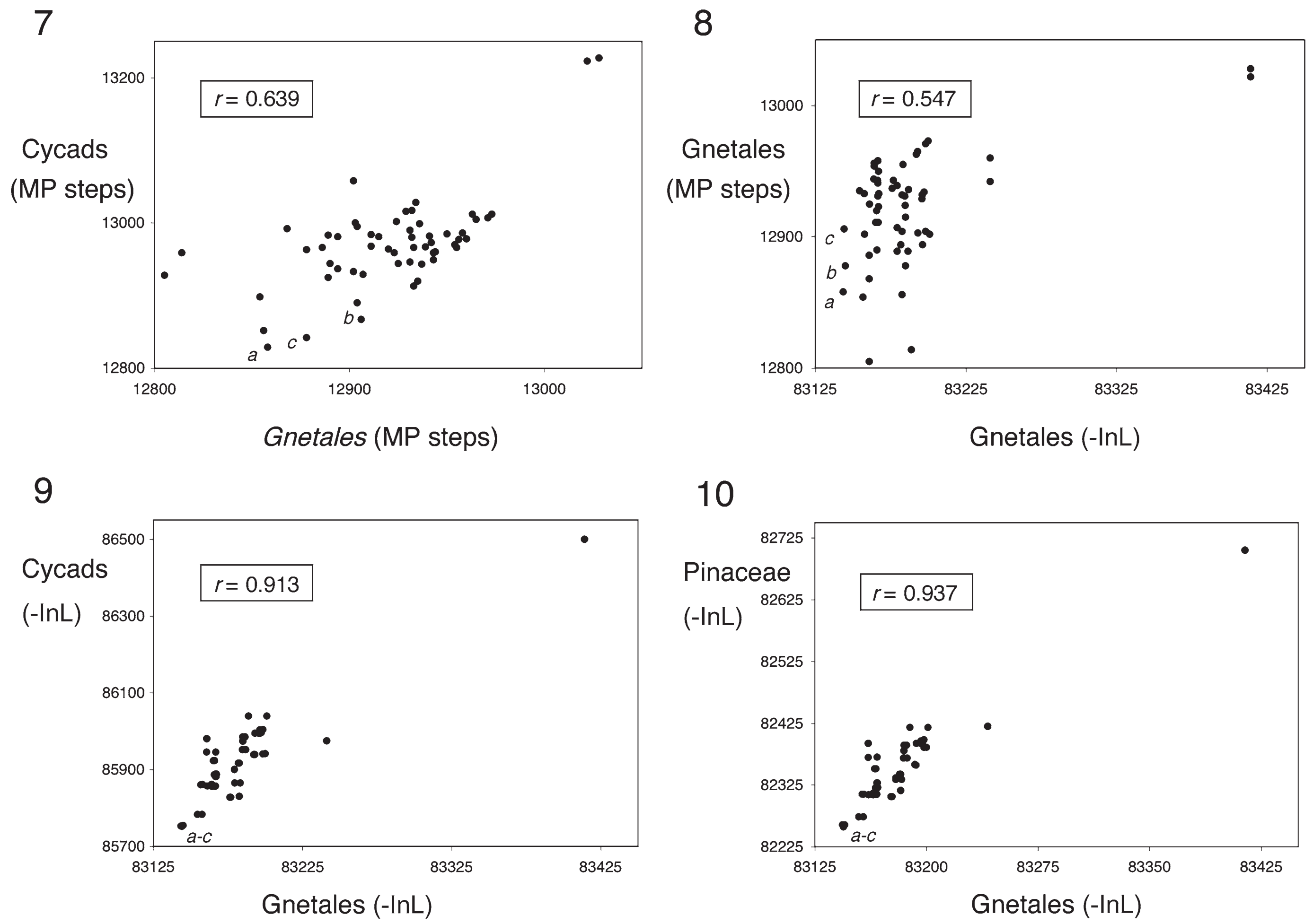
More complex substitutional processes may have occurred that are not properly taken account of in partitioned likelihood analyses (or in corresponding Bayesian analyses), such as heterotachy, a change in substitution rate parameters over time at



Figs. 3–10. Cost of rooting angiosperms along different ingroup branches using *Ginkgo* (Figs. 3–6) or Gnetales (Figs. 7–10), compared to corresponding penalties when alternative outgroups or contrasting optimality criteria (maximum parsimony [MP] vs. maximum likelihood [ML]) are considered. Correlation coefficient, *r*, for each contrasted pair of root cost profiles noted (each is significant; df = 55; $P < 0.01$). Costs are total tree length (–ln L for ML; parsimony steps for MP) when stem branch connecting angiosperms to various gymnosperm outgroup combinations is attached on each of 57 possible ingroup branches on unrooted version of angiosperm tree (Fig. 1). Three optimal or nearly optimal roots marked in each case: *a* = *Amborella* sister to other angiosperms; *b* = *Amborella* + Nymphaeales s.l. sister to other angiosperms; *c* = Only Nymphaeales s.l. sister to other angiosperms. **3.** Contrasted angiosperm root cost profiles for *Ginkgo* (MP) vs. cycads (MP). **4.** Contrasted angiosperm root cost profiles for *Ginkgo* (ML) vs. *Ginkgo* (MP). **5.** Contrasted angiosperm root cost profiles for *Ginkgo* (ML) vs. cycads (ML). **6.** Contrasted angiosperm root cost profiles for *Ginkgo* (ML) vs. Pinaceae (ML).

all sites or subsets of them (Kolaczkowski and Thornton, 2004; Philippe et al., 2005). Currently implemented models for these kinds of substitutional process are likely unable to adequately account for heterotachy as it occurs in real biological data, a problem that will require development of better (more complex) ML models. This is an unsolved problem in phylogenetic analysis (Gruenheit et al., 2008). However, our analyses across most seed plant outgroups (except Gnetales) infer remarkably parallel estimates of the root of angiosperm phylogeny, despite considerable opportunities for the accumulation of misinformative sites and general reduction in signal in each gymnosperm line (i.e., along the outgroup branches that are not shared among them; Fig. 2). All of the seed plant outgroups except Gnetales have virtually identical rooting preferences, particularly in ML analysis (e.g., Fig. 11). Therefore, we speculate that poorly corrected substitutional events along the angiosperm stem branch were not severe enough to grossly mislead inference of the angiosperm root.

Nonetheless, we have also seen that small numbers of nucleotide differences from the site-rate classifications can lead to quite divergent results, at least for Gnetales. This is clearly a problematic outgroup taxon when used on its own, and so may perhaps be disregarded. However, a moderately strong ML bootstrap support for one root position (80–85% support for a root along the branch leading to *Amborella* + Nymphaeales s.l.; Table 1) when using Pinaceae or /Cupressophyta may be indicative of the accumulation of more homoplasy in these divergent outgroups (Table 1), rather than of more signal (see also Table 2, where SH tests find no evidence for a convincing difference in root preference among three nearly optimal roots using these two outgroups). The ML bootstrap results for these two outgroups can be contrasted with those for *Ginkgo* and cycads, which in general are highly comparable to the 36-taxon outgroup combination in terms of their overall performance (Tables 1, 2; Fig. 11). It is possible that *Ginkgo* and cycads have a reduced amount of accumulated homoplasy compared to



7. Contrasted angiosperm root cost profiles for Gnetales (MP) vs. cycads (MP). **8.** Contrasted angiosperm root cost profiles for Gnetales (ML) vs. Gnetales (MP). **9.** Contrasted angiosperm root cost profiles for Gnetales (ML) vs. cycads (ML). **10.** Contrasted angiosperm root cost profiles for Gnetales (ML) vs. Pinaceae (ML).

Pinaceae and /Cupressophyta. If so, this would be consistent with the lower rate of plastid genome evolution for the former two lineages compared to other seed plants (Rai et al., 2003).

None of the outgroup combinations considered here can readily discern among three nearly optimal root placements using SH tests (*Amborella*; *Amborella* plus Nymphaeales; Nymphaeales). Including an extra member of Hydatellaceae here did not help in this regard compared to Saarela et al. (2007), where we only had information for one taxon. This is perhaps because its inclusion did little to break up the relatively long branch leading to a previously sampled member of this family. Clearly, the divergence between these two taxa happened relatively recently. This may reflect a relatively recent origin of the crown clade of Hydatellaceae. We are currently working on a densely sampled phylogeny of the family to address this possibility.

Resolving the root of the crown clade of angiosperms provides the arrow of time for angiosperm phylogeny; the unrooted tree, in contrast, has no reference point for inferring the direction of evolution. Accurate inference of the root node is therefore pivotal to unraveling some of the mysteries surrounding the origin and early radiation of the flowering plants. On the whole, our analyses suggest that the current near-consensus in published studies regarding the root node of angiosperms (i.e., at or very close to the branch leading to *Amborella*) is robust, because most seed plant outgroups have highly congruent signal and the one that does not (Gnetales) is clearly deviant in its behavior within seed plant phylogeny. Nonetheless, definitive resolution on this question may await the addition of more genes from multiple outgroup taxa. We should also be alert to the possibility that better models of DNA sequence evolution may lead to the detection of slight (or perhaps not so slight) systematic biases too subtle to detect using current data and methods. Even a small distortion may be sufficient to nudge the inferred root off course, toward a nearby but incorrect position.

Turning the logic of the current study around, it is also worth considering the possibility that inference of seed plant (or gymnosperm) relationships as a whole may be affected by different samplings of angiosperms in phylogenetic analysis or even by different plausible angiosperm roots. While this hypothesis is worth testing, we predict that it is unlikely to have a major effect so long as angiosperms are sampled sensibly (i.e., by avoiding long ingroup branches like *Arabidopsis*). The crown angiosperms originated relatively recently (perhaps within the last 130 Myr; Magallón and Castillo, 2009, pp. 349–365 in this issue) and radiated very rapidly into the major extant lines, whereas the five gymnosperm lineages considered here (and

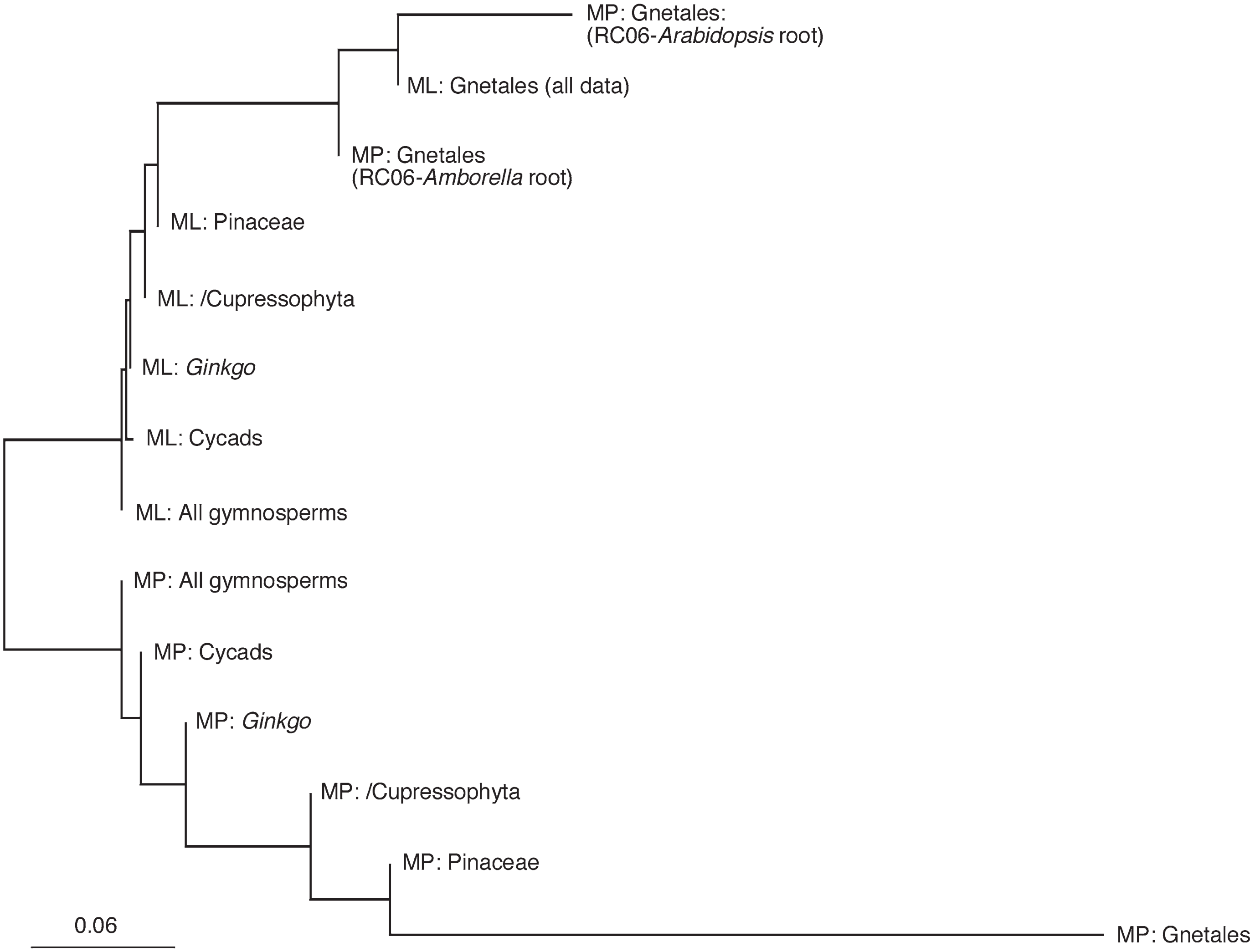


Fig. 11. Neighbor-joining tree summarizing dissimilarity (1 – pairwise correlation coefficient, *r*) of angiosperm root-preference profiles, for different seed-plant outgroup combinations and two phylogenetic optimality criteria (ML, MP; maximum likelihood and parsimony, respectively). MP calculations of root preferences using Gnetales as outgroup were performed using all the data and also using the seven slowest of nine rate classes (RC0–6), as determined in ML classifications based on two extreme root placements for Gnetales (i.e., with *Amborella* sister to all other angiosperms or *Arabidopsis* sister to all other angiosperms).

seed plants as a whole) are all considerably older. Given the different time scales involved in these radiations, there is likely to be only limited opportunity for disruption of seed plant relationships as a whole by the minor but nonnegligible uncertainty that we found concerning the placement of the root node of angiosperm phylogeny.

LITERATURE CITED

Anderson, F. E., and D. L. Swofford. 2004. Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Molecular Phylogenetics and Evolution* 33: 440–451.

Barkman, T. J., G. Chenery, J. R. McNeal, J. Lyons-Weiler, W. J. Ellisens, G. Moore, A. D. Wolfe, and C. W. dePamphilis. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proceedings of the National Academy of Sciences, USA* 97: 13166–13171.

Borsch, T., K. W. Hilu, D. Quandt, V. Wilde, C. Neinhuis, and W. Barthlott. 2003. Noncoding plastid *trnT-trnF* sequences reveal a well resolved phylogeny of basal angiosperms. *Journal of Evolutionary Biology* 16: 558–576.

Burleigh, J. G., and S. Mathews. 2004. Phylogenetic signal in nucleotide data from seed plants: Implications for resolving the seed plant tree of life. *American Journal of Botany* 91: 1599–1613.

Cai, Z., C. Penaflor, J. V. Kuehl, J. Leebens-Mack, J. E. Carlson, C. W. dePamphilis, J. L. Boore, and R. K. Jansen. 2006. Complete plastid genome sequences of *Drimys, Liriodendron*, and *Piper*: Implications for the phylogenetic relationships of magnoliids. *BMC Evolutionary Biology* 6: 77 [online, doi:10.1186/1471-2148-6-77].

Cantino, P. D., J. A. Doyle, S. W. Graham, W. S. Judd, R. G. Olmstead, D. E. Soltis, P. S. Soltis, and M. J. Donoghue. 2007. Towards a phylogenetic nomenclature of *Tracheophyta. Taxon* 56: 822–846.

Crane, P. R., P. Herendeen, and E. M. Friis. 2004. Fossils and plant phylogeny. *American Journal of Botany* 91: 1683–1699.

Doyle, J. A. 2006. Seed ferns and the origin of the angiosperms. *Journal of the Torrey Botanical Society* 133: 169–209.

DOYLE, J. A. 2008. Integrating molecular phylogenetic and paleobotanical evidence on origin of the flower. *International Journal of Plant Sciences* 169: 816–843.

DOYLE, J. A., AND P. K. ENDRESS. 2000. Morphological phylogenetic analysis of basal angiosperms: comparison and combination with molecular data. *International Journal of Plant Sciences* 161 (Supplement 6): S121–S153.

ENDRESS, P. K. 2008. Perianth biology in the basal grade of extant angiosperms. *International Journal of Plant Sciences* 169: 844–862.

ENDRESS, P. K., AND J. A. DOYLE. 2009. Reconstructing the ancestral angiosperm flower and its initial specializations. *American Journal of Botany* 96: 22–66.

FEILD, T. S., AND N. C. ARENS. 2007. The ecophysiology of early angiosperms. *Plant, Cell & Environment* 30: 291–309.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401–410.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39: 783–791.

FELSENSTEIN, J. 2007. PHYLIP: Phylogenetic inference package, version 3.67. Department of Genome Sciences and Department of Biology, University of Washington, Seattle, Washington, USA.

FRIEDMAN, W. E. 2008. Hydatellaceae are water lilies with gymnospermous tendencies. *Nature* 453: 94–97.

FRIIS, E. M., AND P. CRANE. 2007. New home for tiny aquatics. *Nature* 446: 269–270.

FRIIS, E. M., K. R. PEDERSEN, AND P. R. CRANE. 2006. Cretaceous angiosperm flowers: Innovation and evolution in plant reproduction. *Palaeogeography, Palaeoclimatology, Palaeoecology* 232: 251–293.

GAUT, B. S., S. V. MUSE, W. D. CLARK, AND M. T. CLEGG. 1992. Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. *Journal of Molecular Evolution* 35: 292–303.

GOREMYKIN, V. V., K. I. HIRSCH-ERNST, S. WÖLFL, AND F. H. HELLWIG. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Molecular Biology and Evolution* 20: 1499–1505.

GOREMYKIN, V. V., K. I. HIRSCH-ERNST, S. WÖLFL, AND F. H. HELLWIG. 2004. The chloroplast genome of *Nymphaea alba*: Whole-genome analyses and the problem of identifying the most basal angiosperm. *Molecular Biology and Evolution* 21: 1445–1454.

GOREMYKIN, V. V., B. HOLLAND, K. I. HIRSCH-ERNST, AND F. H. HELLWIG. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Molecular Biology and Evolution* 22: 1813–1822.

GRAHAM, S. W., AND R. G. OLMSTEAD. 2000a. Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Current Genetics* 37: 183–188.

GRAHAM, S. W., AND R. G. OLMSTEAD. 2000b. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. *American Journal of Botany* 87: 1712–1730.

GRAHAM, S. W., R. G. OLMSTEAD, AND S. C. H. BARRETT. 2002. Rooting phylogenetic trees with distant outgroups: a case study from the commelinoid monocots. *Molecular Biology and Evolution* 19: 1769–1781.

GRAHAM, S. W., P. A. REEVES, A. C. E. BURNS, AND R. G. OLMSTEAD. 2000. Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *International Journal of Plant Sciences* 161 (Supplement 6): S83–S96.

GRAHAM, S. W., J. M. ZGURSKI, M. A. MCPHERSON, D. M. CHERNIAWSKY, J. M. SAARELA, E. F. C. HORNE, S. Y. SMITH, et al. 2006. Robust inference of monocot deep phylogeny using a expanded multigene plastid data set. *In* J. T. Columbus, E. A. Friar, J. M. Porter, L. M. Prince, and M. G. Simpson [eds.], Monocots: Comparative biology and evolution (excluding Poales), 3–21. Rancho Santa Ana Botanic Garden, Claremont, California, USA.

GRUENHEIT, N., P. J. LOCKHART, M. STEEL, AND W. MARTIN. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. *Molecular Biology and Evolution* 25: 1512–1520.

GUINDON, S., AND O. GASCUEL. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52: 696–704.

HANSEN, D. R., S. G. DASTIDAR, Z. CAI, C. PENAFLOR, J. V. KUEHL, J. L. BOORE, AND R. K. JANSEN. 2007. Phylogenetic and evolutionary implications of complete chloroplast genome sequences of four early-diverging angiosperms: *Buxus* (Buxaceae), *Chloranthus* (Chloranthaceae), *Dioscorea* (Dioscoreaceae), *Illicium* (Schisandraceae). *Molecular Phylogenetics and Evolution* 45: 547–563.

HEDTKE, S. M., T. M. TOWNSEND, AND D. M. HILLIS. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology* 55: 522–529.

HENDY, M. D., AND D. PENNY. 1989. A framework for quantitative study of evolutionary trees. *Systematic Zoology* 38: 297–309.

HILLIS, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* 47: 3–8.

HILLIS, D. M., D. D. POLLOCK, J. A. MCGUIRE, AND D. J. ZWICKL. 2003. Is sparse taxon sampling a problem for phylogenetic inference? *Systematic Biology* 52: 124–126.

HILU, K. W., T. BORSCH, K. MÜLLER, D. E. SOLTIS, P. S. SOLTIS, V. SAVOLAINEN, M. W. CHASE, et al. 2003. Angiosperm phylogeny based on *matK* sequence information. *American Journal of Botany* 90: 1758–1776.

HOLLAND, B. R., D. PENNY, AND M. D. HENDY. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—A simulation study. *Systematic Biology* 52: 229–238.

JANSEN, R. K., Z. CAI, L. A. RAUBESON, H. DANIELL, C. W. DEPAMPHILIS, J. LEEBENS-MACK, K. F. MÜLLER, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proceedings of the National Academy of Sciences, USA* 104: 19369–19374.

JUDD, W. S., C. S. CAMPBELL, E. A. KELLOGG, P. S. STEVENS, AND M. J. DONOGHUE. 2008. Plant systematics: A phylogenetic approach. Sinauer, Sunderland, Massachusetts, USA.

KOLACZKOWSKI, B., AND J. W. THORNTON. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431: 980–984.

KOSAKOVSKY POND, S. L., S. D. W. FROST, AND S. V. MUSE. 2005. HyPhy: Hypothesis testing using phylogenies. *Bioinformatics (Oxford, England)* 21: 676–679.

LEEBENS-MACK, J., L. A. RAUBESON, L. CUI, J. V. KUEHL, M. H. FOURCADE, T. W. CHUMLEY, J. L. BOORE, R. K. JANSEN, AND C. W. DEPAMPHILIS. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. *Molecular Biology and Evolution* 22: 1948–1963.

LOCKHART, P. J., AND D. PENNY. 2005. The place of *Amborella* within the radiation of angiosperms. *Trends in Plant Science* 10: 201–202.

MADDISON, D. R., AND W. P. MADDISON. 2001. MacClade 4: Analysis of phylogeny and character evolution, version 4.03. Sinauer, Sunderland, Massachusetts, USA.

MAGALLÓN, S. AND A. CASTILLO. 2009. The roots of angiosperm diversity. *American Journal of Botany* 96: 349–365.

MARDANOV, A. V., N. V. RAVIN, B. B. KUZNETSOV, T. H. SAMIGULLIN, A. S. ANTONOV, T. V. KOLGANOVA, AND K. G. SKYABIN. 2008. Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. *Journal of Molecular Evolution* 66: 555–564.

MARTIN, W., O. DEUSCH, N. STAWSKI, N. GRÜNHEIT, AND V. GOREMYKIN. 2005. Chloroplast genome phylogenetics: Why we need independent approaches to plant molecular evolution. *Trends in Plant Science* 10: 203–209.

MATHEWS, S. 2009. Phylogenetic relationships among seed plants: Persistent questions and the limits of DNA sequence data. *American Journal of Botany* 96: 228–236.

MATHEWS, S., AND M. J. DONOGHUE. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286: 947–950.

MATHEWS, S., AND M. J. DONOGHUE. 2000. Basal angiosperm phylogeny inferred from duplicate phytochromes A and C. *International Journal of Plant Sciences* 161 (Supplement 6): S41–S55.

Matsen, F. A., and M. Steel. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology* 56: 767–775.

Moore, M. J., C. D. Bell, P. S. Soltis, and D. E. Soltis. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. *Proceedings of the National Academy of Sciences, USA* 104: 19363–19368.

Müller, K. F., T. Borsch, and K. W. Hilu. 2006. Phylogenetic utility of rapidly evolving DNA at high taxonomical levels: contrasting *matK, trnT-F*, and *rbcL* in basal angiosperms. *Molecular Phylogenetics and Evolution* 41: 99–117.

Murdock, A. G. 2008. Phylogeny of marattioid ferns (Marattiaceae): Inferring a root in the absence of a closely related outgroup. *American Journal of Botany* 95: 626–641.

Nickerson, J., and G. Drouin. 2004. The sequence of the largest subunit of RNA polymerase II is a useful marker for inferring seed plant phylogeny. *Molecular Phylogenetics and Evolution* 31: 403–415.

Parkinson, C. L., K. L. Adams, and J. D. Palmer. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. *Current Biology* 9: 1485–1488.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology* 5: 50 [online, doi:10.1186/1471-2148-5-50].

Posada, D., and K. A. Crandall. 1998. MODELTEST: Testing the model of DNA substitution. *Bioinformatics (Oxford, England)* 14: 817–818.

Qiu, Y.-L., O. Dombrovska, J. Lee, L. Li, B. A. Whitlock, F. Bernasconi-Quadroni, J. S. Rest, et al. 2005. Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. *International Journal of Plant Sciences* 166: 815–842.

Qiu, Y.-L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: Evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404–407.

Qiu, Y.-L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. 2000. Phylogeny of basal angiosperms: Analyses of five genes from three genomes. *International Journal of Plant Sciences* 161 (Supplement 6): S3–S27.

Qiu, Y.-L., J. Lee, B. A. Whitlock, F. Bernasconi-Quadroni, and O. Dombrovska. 2001. Was the ANITA rooting of the angiosperm phylogeny affected by long-branch attraction? *Molecular Biology and Evolution* 18: 1745–1753.

Qiu, Y.-L., L. Li, T. A. Hendry, R. Li, D. W. Taylor, M. J. Issa, A. J. Ronen, M. L. Vekaria, and A. M. White. 2006. Reconstructing the basal angiosperm phylogeny: Evaluating information content of mitochondrial genes. *Taxon* 55: 837–856.

Rai, H. S., H. E. O'Brien, P. A. Reeves, R. G. Olmstead, and S. W. Graham. 2003. Inference of higher-order relationships in the cycads from a large chloroplast data set. *Molecular Phylogenetics and Evolution* 29: 350–359.

Rai, H. S., P. A. Reeves, R. Peakall, R. G. Olmstead, and S. W. Graham. 2008. Inference of higher-order conifer relationships from a multi-locus plastid data set. *Botany* 86: 658–669.

Raven, P. H., R. F. Evert, and S. E. Eichhorn. 2005. Biology of plants, 7th ed. Freeman, New York, New York, USA.

Remizowa, M. V., D. D. Sokoloff, T. D. Macfarlane, S. Y. Yadav, C. J. Prychid, and P. J. Rudall. 2008. Comparative pollen morphology in the early-divergent angiosperm family Hydatellaceae reveals variation at the infraspecific level. *Grana* 47: 81–100.

Rodríguez-Ezpeleta, N., H. Brinckmann, G. Burger, A. J. Roger, M. W. Gray, H. Philippe, and B. F. Lang. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. *Current Biology* 17: 1420–1425.

Rothwell, G. W., W. L. Crepet, and R. A. Stockey. 2009. Is the anthophyte hypothesis alive and well? New evidence from the reproductive structures of Bennettitales. *American Journal of Botany* 96: 296–322.

Rothwell, G. W., and R. A. Stockey. 2002. Anatomically preserved *Cycadeoidea* (Cycadeoidaceae), with a reevaluation of systematic characters for the seed cones of Bennettitales. *American Journal of Botany* 89: 1447–1458.

Rudall, P. J., M. V. Remizowa, A. S. Beer, E. Bradshaw, D. S. Stevenson, T. D. Macfarlane, R. E. Tuckett, S. R. Yadav, and D. D. Sokoloff. 2008. Comparative ovule and megagametophyte development in Hydatellaceae and water lilies reveal a mosaic of features among the earliest angiosperms. *Annals of Botany* 101: 941–956.

Rudall, P. J., D. D. Sokoloff, M. V. Remizowa, J. C. Conran, J. I. Davis, T. D. Macfarlane, and D. W. Stevenson. 2007. Morphology of Hydatellaceae, an anomalous aquatic family recently recognized as an early-divergent angiosperm lineage. *American Journal of Botany* 94: 1073–1092.

Saarela, J. M., H. S. Rai, J. A. Doyle, P. K. Endress, S. Mathews, A. D. Marchant, B. G. Briggs, and S. W. Graham. 2007. Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. *Nature* 446: 312–315.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. Sher Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. *Molecular Biology and Evolution* 17: 782–797.

Shavit, L., D. Penny, M. D. Hendy, and B. R. Holland. 2007. The problem of rooting rapid radiations. *Molecular Biology and Evolution* 24: 2400–2411.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution* 16: 1114–1116.

Sokoloff, D. D., M. V. Remizowa, T. D. Macfarlane, R. E. Tuckett, M. M. Ramsay, A. S. Beer, S. R. Yadav, and P. J. Rudall. 2008. Seedling diversity in Hydatellaceae: Implications for the evolution of angiosperm cotyledons. *Annals of Botany* 101: 153–164.

Soltis, D. E., V. A. Albert, J. Leebens-Mack, J. D. Palmer, R. A. Wing, C. W. dePamphilis, H. Ma, et al. 2008. The *Amborella* genome: An evolutionary reference for plant biology. *Genome Biology* 9: 402 [online, doi:10.1186/gb-2008-9-3-402].

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y.-L. Qiu, M. W. Chase, J. S. Farris, et al. 2004. Genome-scale data, angiosperm relationships, and 'ending incongruence': A cautionary tale in phylogenetics. *Trends in Plant Science* 9: 477–483.

Soltis, D. E., and P. S. Soltis. 2004. *Amborella* not a "basal angiosperm"? Not so fast. *American Journal of Botany* 91: 997–1001.

Soltis, D. E., P. S. Soltis, M. W. Chase, M. E. Mort, D. C. Albach, M. Zanis, V. Savolainen, et al. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133: 381–461.

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

Stefanović, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evolutionary Biology* 4: 35 [online, doi:10.1186/1471-2148-4-35].

Swofford, D. L. 2002. PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sinauer, Sunderland, Massachusetts, USA.

Zanis, M. J., D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. 2002. The root of the angiosperms revisited. *Proceedings of the National Academy of Sciences, USA* 99: 6848–6853.

Zgurski, J. M., H. S. Rai, Q. M. Fai, D. J. Bogler, J. Francisco-Ortega, and S. W. Graham. 2008. How well do we understand the overall backbone of cycad phylogeny? New insights from a large, multigene plastid data set. *Molecular Phylogenetics and Evolution* 47: 1232–1237.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly decreases phylogenetic error. *Systematic Biology* 51: 588–598.