American Journal of Botany 91(10): 1599–1613. 2004.

PHYLOGENETIC SIGNAL IN NUCLEOTIDE DATA FROM SEED PLANTS: IMPLICATIONS FOR RESOLVING THE SEED PLANT TREE OF LIFE[1]

J. GORDON BURLEIGH[2,4] AND SARAH MATHEWS[3]

[2]Section of Evolution and Ecology, University of California, Davis, California 95616 USA; and [3]Arnold Arboretum of Harvard University, Cambridge, Massachusetts 02138 USA

Effects of taxonomic sampling and conflicting signal on the inference of seed plant trees supported in previous molecular analyses were explored using 13 single-locus data sets. Changing the number of taxa in single-locus analyses had limited effects on log likelihood differences between the gnepine (Gnetales plus Pinaceae) and gnetifer (Gnetales plus conifers) trees. Distinguishing among these trees also was little affected by the use of different substitution parameters. The 13-locus combined data set was partitioned into nine classes based on substitution rates. Sites evolving at intermediate rates had the best likelihood and parsimony scores on gnepine trees, and those evolving at the fastest rates had the best parsimony scores on Gnetales-sister trees (Gnetales plus other seed plants). When the fastest evolving sites were excluded from parsimony analyses, well-supported gnepine trees were inferred from the combined data and from each genomic partition. When all sites were included, Gnetales-sister trees were inferred from the combined data, whereas a different tree was inferred from each genomic partition. Maximum likelihood trees from the combined data and from each genomic partition were well-supported gnepine trees. A preliminary stratigraphic test highlights the poor fit of Gnetales-sister trees to the fossil data.

Key words: Gnetales; multilocus analyses; phylogenetic signal; rate class; seed plant phylogeny; taxonomic sampling.

Phylogenetic relationships among the five extant lines of seed plants remain controversial despite the recent accumulation of molecular data sets to address the question (reviewed in Magallón and Sanderson, 2002). These five lines comprise cycads, ginkgos, conifers, Gnetales, and angiosperms. Stratigraphic evidence places the origin of cycads, ginkgos, and conifers in the Paleozoic, with Gnetales and modern conifer families appearing in the Triassic to Jurassic, and angiosperms later in the Mesozoic (Stewart and Rothwell, 1993; Crane, 1996). From the Permian through the late Jurassic many seed plant lineages went extinct, including lyginopterids, medullosans, Callistophytaceae, glossopterids, Cordaitales, and Voltziales (Stewart and Rothwell, 1993), and their relationships with extant groups remain poorly characterized. Additionally, during the Cretaceous and Tertiary, the diversity of all surviving seed plant lines except angiosperms decreased (Knoll, 1984; Crane, 1987). Thus, as is common in studies of deep divergences, taxonomic diversity is incompletely captured in molecular data sets. Moreover, extant lines vary markedly with respect to levels of current diversity (cf., angiosperms with ~260 000 species and ginkgos with one species) and rates of divergence. In Gnetales, low diversity (70 species in three genera, *Ephedra*, *Gnetum*, and *Welwitschia*) is combined with high rates of divergence. Not surprisingly, the position of Gnetales has been one of the more enigmatic questions in studies of seed plant phylogeny, with molecular data sets strongly supporting alternative hypotheses (Fig. 1).

Prior to the use of cladistic methods, competing hypotheses united Gnetales with conifers (Bailey, 1944; Eames, 1952; Takhtajan, 1969; Bierhorst, 1971; Doyle, 1978) or with angiosperms (Arber and Parkin, 1907, 1908; Wettstein, 1907). Chamberlain (1935) included Gnetales in his Coniferophytes along with conifers, ginkgos, and Cordaitales but considered their placement problematic and did not rule out a relationship with angiosperms. Cronquist (1968) and Thorne (1976) rejected a relationship with angiosperms, but did not strongly advocate an alternative position for Gnetales. Nonetheless, a series of cladistic analyses of morphological data united Gnetales with angiosperms (Parenti, 1980; Crane, 1985; Doyle and Donoghue, 1986; Loconte and Stevenson, 1990; Nixon et al., 1994; Rothwell and Serbet, 1994; Doyle, 1996, 1998b), seeming to confirm the views of Arber and Parkin (1907, 1908) and Wettstein (1907). In most cladistic analyses of morphological characters that included fossil taxa, angiosperms, Gnetales, Bennettitales, and *Pentoxylon* formed a clade (Crane, 1985; Doyle and Donoghue, 1986, 1992; Nixon et al., 1994; Rothwell and Serbet, 1994). The term ''anthophytes'' was used for this clade because the aggregations of sporophylls in each line were interpreted as flower-like structures (Doyle and Donoghue, 1987). Doyle (1996) later found a glossophyte clade that nested *Caytonia* within the anthophytes and placed glossopterids as their sister clade. Gnetales were sister to angiosperms in the trees of Crane (1985) and Rothwell and Serbet (1994) but not in the trees of Doyle (1996) or Doyle and Donoghue (1986, 1992). Nixon et al. (1994) found trees with angiosperms nested within Gnetales. Nonetheless, the morphological analyses were consistent in supporting the anthophyte hypothesis (Doyle, 1998a).

This result was challenged when early analyses of molecular data placed Gnetales as sister to all remaining seed plants (Hamby and Zimmer, 1992; Albert et al., 1994) in ''Gnetales-sister'' trees or as sister to all other extant gymnosperms (Fig. 1; Hasebe et al., 1992; Goremykin et al., 1996). More recently, trees with Gnetales sister to all other extant gymnosperms also

[1] Manuscript received 26 January 2004; revision accepted 17 June 2004.

The authors gratefully acknowledge the assistance of Amy Driskell, Rosita Scherson, and Lisa Thurston. Jim Doyle, Mike Sanderson, Sean Graham, and two anonymous reviewers provided helpful comments on this paper. Also, the editors of this volume, Mark Chase, Jeff Palmer, and Doug Soltis, made valuable comments on the manuscript. This material is based upon work supported by the National Science Foundation under grant numbers 1053164 (Burleigh) and 0196150 (Mathews).

[4] E-mail: jgburleigh@ucdavis.edu.

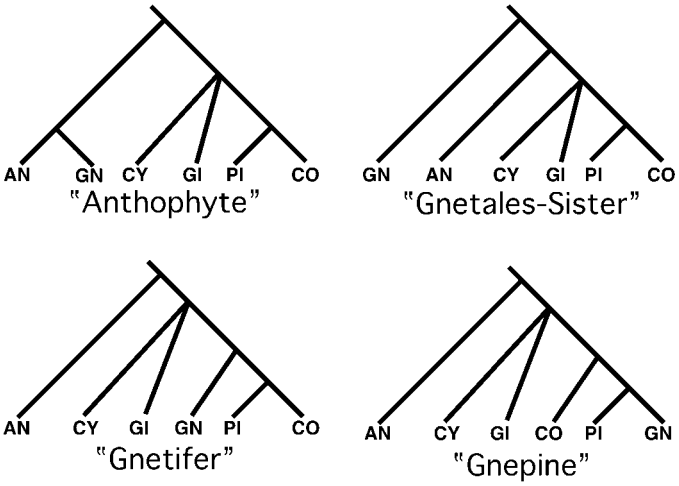


Fig. 1. Four major hypotheses of relationships among extant seed plant lineages. AN = angiosperms; CY = cycads; GI = *Ginkgo*; GN = Gnetales; PI = Pinaceae; CO = non-Pinaceae conifers.

were recovered in parsimony analyses of plastid *rpoC1* (Samigullin et al., 1999), *AGL*6 and *AGL-like* genes (Winter et al., 1999; Becker and Theissen, 2003), *Floricaula*/*LEAFY* (Frohlich and Parker, 2000), and phytochromes (Mathews and Donoghue, 2002; Schmidt and Schneider-Poetsch, 2002). Trees that placed Gnetales with angiosperms, consistent with an anthophyte concept, were rare. Hamby and Zimmer (1992) found that neighbor joining analyses of their ribosomal RNA data united Gnetales with angiosperms and that a parsimony tree with this clade was just one step longer than the Gnetales-sister tree. Parsimony analysis of 26S ribosomal DNA (rDNA) also united Gnetales and angiosperms but with bootstrap support below 50% and a Bremer support value of one (Stefanovic et al., 1998). Rydin et al. (2002) also obtained anthophyte trees from unweighted parsimony analyses of a 26S rDNA and a combined 26S rDNA and 18S rDNA data set. The lack of support for the anthophyte hypothesis is consistent with evidence that it would be difficult to infer from *psaA* or *psbB* using parsimony even if it were true (e.g., Sanderson et al., 2000).

Many subsequent analyses of molecular data provided support for the previous suggestion that Gnetales were related to conifers (Bailey, 1944; Eames, 1952; Bierhorst, 1971; Doyle, 1978). Analyses of 18S rDNA contradicted the trees of Hamby and Zimmer (1992), uniting Gnetales with a monophyletic conifer clade in ''gnetifer'' trees (Fig. 1; Chaw et al., 1997, 2000; Bowe et al., 2000; Rydin et al., 2002; Soltis et al., 2002), and several gene trees united Pinaceae (other conifers were not sampled) with Gnetales (Goremykin et al., 1996; Hansen et al., 1999; Samigullin et al., 1999; Winter et al., 1999; Antonov et al., 2000; Nickerson and Drouin, 2004). A series of multilocus analyses further suggested a close relationship between conifers and Gnetales but indicated that conifers were paraphyletic, uniting Gnetales with Pinaceae in ''gnepine'' trees (Fig. 1; Qiu et al., 1999; Bowe et al., 2000; Chaw et al., 2000; Nickrent et al., 2000; Gugerli et al., 2001). Gnepine trees also were obtained when 19 exemplars were sampled for eight loci representing all three genomes (Soltis et al., 2002). In contrast, analyses of a two-genome data set representing just nuclear and plastid genes but from many more taxa yielded highly supported Gnetales-sister trees in parsimony analyses (Rydin et al., 2002; contra Soltis et al., 2002, p. 1679). Parsimony analyses of all sites in a ~13.5-kb multilocus plastid data set also gave a Gnetales-sister tree (Rai et al., 2003). The investigation of rooted and unrooted trees by Bowe et al. (2000) partially addressed the possibility that long-branch effects, introduced by the inclusion of divergent outgroups, were influencing the topology of seed plant trees. If their unrooted trees were rooted at Gnetales, Pinaceae or all conifers would diverge next from the remaining seed plants, conflicting with the topology of Gnetales-sister trees inferred in analyses that included outgroups, in which the next split was between angiosperms and the remaining seed plants (Hamby and Zimmer, 1992; Albert et al., 1994; Rydin et al., 2002; Rai et al., 2003). This suggests that inclusion of divergent outgroups might have influenced the rooting of Gnetales-sister trees.

In addition, several analyses revealed conflict and bias within molecular data sets. Third codon positions of two plastid genes *psaA* and *psbB* supported a Gnetales-sister tree, whereas first and second positions supported a gnepine tree; bias was detected at third positions in *psaA* and *psbB* and in first and second positions of *psbB* (Sanderson et al., 2000; Magallón and Sanderson, 2002). Additionally, hypothesis testing of the same *psaA* and *psbB* data sets using models that account for heterogeneity in the process of evolution among codon positions supported a gnepine hypothesis (Aris-Brosou, 2003). Rydin et al. (2002) found that two plastid genes supported a Gnetales-sister tree if transitions were included but supported gnepine trees when transitions were excluded. Analyses of the plastid sequences in the combined data set of Soltis et al. (2002) yielded results similar to those of Sanderson et al. (2000) and Magallón and Sanderson (2002); maximum parsimony (MP) analyses of all codon positions or of just third codon positions gave Gnetales-sister trees, whereas maximum likelihood (ML) and MP analyses of first and second codon positions gave gnepine trees (Soltis et al., 2002). A series of analyses also revealed that the use of different optimality criteria gave conflicting but well-supported trees (Samigullin et al., 1999; Bowe et al., 2000; Frohlich and Parker, 2000; Mathews and Donoghue, 2002; see detailed summary in Magallón and Sanderson, 2002). Notably, the trees differed markedly in several cases, with MP analyses placing Gnetales as sister to all other gymnosperms and ML or Bayesian analyses placing them as sister to Pinaceae (Samigullin, 1999; Frohlich and Parker, 2000; Mathews and Donoghue, 2002; Rydin and Källersjö, 2002). Finally, Rydin and Källersjö (2002) found that the *rbcL* data sets gave different results, depending on character weighting, method of analysis, and taxonomic sampling.

In summary, results of molecular analyses have highlighted the difficulty of resolving seed plant relationships, and the three hypotheses that receive the most support in molecular analyses, the gnepine, gnetifer, and Gnetales-sister trees, are not completely consistent with all the available data. Thus, it would be productive to carefully examine the available molecular data for further insight into the causes of ambiguity. Previous studies have suggested that taxonomic sampling may be partly responsible for inconsistent results (e.g., Rydin and Källersjö, 2002; Soltis et al., 2002), and this is a plausible explanation for the disparity among published trees. For example, the analyses of Qiu et al. (1999), Bowe et al. (2000), Chaw et al. (2000), Gugerli et al. (2001), Soltis et al. (2002), and Rai et al. (2003) included 8, 17, 19, 17, 11, and 16 gymnosperms, respectively, whereas the analyses of Rydin et al. (2002) included 69 gymnosperms. Several studies also have

revealed that heterogeneity in rates of nucleotide evolution across sites could be a confounding factor (e.g., Chaw et al., 2000; Sanderson et al., 2000; Magallón and Sanderson, 2002; Rydin et al., 2002; Soltis et al., 2002; Aris-Brosou, 2003). Other studies have shown that the choice of optimality criterion has an effect (e.g., Hamby and Zimmer, 1992; Hasebe et al., 1992; Bowe et al., 2000; Chaw et al., 2000; Frohlich and Parker, 2000; Magallón and Sanderson, 2002; Mathews and Donoghue, 2002). We conducted a series of analyses using data from 13 loci to further explore the effects of these factors on inference of seed plant phylogeny.

MATERIALS AND METHODS

Genes and taxonomic sampling—We sampled data from GenBank (http://www.ncbi.nlm.nih.gov) in a way that maximized both the taxonomic sampling among gymnosperms and the number of loci sampled. We sampled loci that, at a minimum, had sequence data from at least one taxon from each of the major clades of seed plants as well as representative conifers, including members of Araucariaceae, Cupressaceae, Pinaceae, Podocarpaceae, and Taxaceae. Only Cephalotaxaceae, Phyllocladaceae, and the monotypic Sciadopityaceae were not included. Molecular data indicate that Cephalotaxaceae and Phyllocladaceae are members of Taxaceae and Podocarpaceae, respectively (Quinn et al., 2002). Additionally, the locus had to be alignable across all seed plants. We augmented phytochrome data in GenBank with unpublished sequences, focusing on two loci, *PHYP*, an unambiguous homolog of angiosperm *PHYB*, and *PHYN*, a putative homolog of angiosperm *PHYA* (Sharrock and Mathews, in press). In total, we sampled 13 loci, including four nuclear loci (18S rDNA, 26S rDNA, *PHYP/B*, and *PHYN/A*), five plastid loci (*atpB, matK, psaA, psbB*, and *rbcL*) and four mitochondrial loci (*atpA, coxI, mtSSU*, and *nad*5). For each of the 13 loci, we searched GenBank for all available gymnosperm sequences. We also included *Equisetum* and selected ferns as outgroups (Pryer et al., 2001). For the angiosperms, we selected sequences from four basal lineages: Amborellaceae, Nymphaeales, Austrobaileyales, and Chloranthaceae (e.g., Mathews and Donoghue, 1999; Parkinson et al., 1999; Qiu et al., 1999; Doyle and Endress, 2000). The single-locus data sets that contain all available gymnosperm sequences are referred to as the *complete taxon sampling* data sets. Each complete taxon sampling data set was aligned using AVID (Bray et al., 2003; http://baboon.math.berkeley.edu/mavid/), and then manually adjusted. The alignments of 26S rDNA, *matK*, and *mtSSU* had large regions that could not be easily aligned, and these regions were excluded from analyses. We removed all *matK* outgroup sequences because they were extremely difficult to align with the seed plant sequences. The *PHYP/B* and *PHYN/A* alignments contain two indel regions of approximately 20 nucleotides each that also were excluded. The accession tables from the complete sampling data sets are archived in the Appendix (see Supplemental Data accompanying online version of this article).

To combine the single-locus data sets, we identified a set of 31 exemplar genera, including 21 gymnosperm genera, for which sequences were available from at least 10 of the 13 loci (see Appendix in Supplemental Data). In some of the complete taxon sampling data sets, certain genera were represented by multiple species. In these cases, all but one species per genus were removed, and the resulting single-locus data sets are referred to as the *limited taxon sampling* data sets. The limited taxon sampling data sets were combined to build multilocus *combined data sets.* This sometimes required merging sequences from different species to represent a single genus (see Appendix). We built combined data sets that included all loci, only nuclear loci, only plastid loci, and only mitochondrial loci. The combined data set of all loci contained 31 genera and is 18 906 nucleotides long. The nucleotide sequence sampled for a single taxon ranged from 7785 nucleotides to 18 042 nucleotides, and the median for a taxon was 12 760 nucleotides. In the combined data set, 32.5% (31.1% nuclear, 23.9% plastid, and 47.0% mitochondrial data) of the cells are coded as missing data. Though this is a large percentage of missing data, the number of complete characters, which may have a greater effect on phylogenetic accuracy (Wiens, 2003), is also large.

Phylogenetic analyses—We performed ML phylogenetic analyses on each of the complete taxon sampling, limited taxon sampling, and combined data sets using PAUP* (Swofford, 2002). In each analysis, we used a general reversible model (e.g., Tavaré, 1986) with rate variation among sites estimated using a discrete gamma distribution with four categories (Yang, 1994) and a separate category for the percentage of invariable sites. To improve the computational tractability of analyses that implement this complex model, we estimated all substitution parameters from a parsimony tree and used these estimates in the ML tree search. We found that the estimates of substitution parameters varied little among major seed plant hypotheses (not shown), consistent with the observation that substitution parameters are often similar over a wide range of trees (e.g., Yang et al., 1994; Sullivan et al., 1996). Our ML analyses used a heuristic tree search algorithm consisting of one run of random sequence addition and TBR branch swapping (Swofford et al., 1996). In the case of three of the largest trees (the complete taxon sampling trees from 18S rDNA, *rbcL*, and *matK*), the running time of the branch swapping was limited to 1 week. We performed 100 replicates of nonparametric bootstrapping to assess the confidence in the ML topology (Felsenstein, 1985). The tree search for the bootstrap replicates was identical to the initial tree search. However, the single-locus, complete taxon sampling data sets for 18S rDNA, 26S rDNA, *atpB, matK*, and *rbcL* were too large to bootstrap with the available computational resources.

We also performed MP analyss on the combined data sets. The MP analyses used a heuristic tree search with 1000 random addition sequence replicates and TBR branch swapping. To assess support, we performed 1000 bootstrap replicates, each with 10 random sequence addition replicates.

Effects of taxonomic sampling and substitution parameters on phylogenetic inference—Although the effect of taxonomic sampling on phylogenetic inference remains controversial (e.g., Rosenberg and Kumar, 2001; Pollock et al., 2002), it may influence seed plant analyses (Magallón and Sanderson, 2002; Rydin and Källersjö, 2002; Soltis et al., 2002). Thus, while the number of extinctions within seed plants limits our ability to explore the effects of taxonomic sampling on the inference of seed plant phylogeny, it is important to explore its effects on the inference of trees from molecular data, particularly given the small number of gymnosperms included in most molecular analyses of seed plant data. We examined the effect of taxonomic sampling on ML analyses by comparing the seed plant trees reconstructed from the complete taxon sampling data sets, which included up to 361 taxa, with trees reconstructed from the limited taxon sampling data sets, which contained at most 30 sequences. Specifically, for each locus we compared ln likelihood differences (δ) between gnepine and gnetifer trees to see if δ changed with changes in taxonomic sampling. These tests focused on how taxonomic sampling might affect the ability to distinguish between gnetifer and gnepine hypotheses. The taxonomic sampling for *nad5* was sparse, and thus, we excluded it from this test.

We also examined how robust the phylogenies from the limited taxon sampling data sets were to changes in the estimated parameter values. For example, estimates of rate variation may be more accurate when taxonomic sampling is greater (Sullivan et al., 1999). In particular, small data sets may overestimate the percentage of invariable sites and the shape parameter (α) of the gamma distribution that describes rate variation among sites (Sullivan et al., 1999). We repeated the ML analyses (unconstrained and constrained to gnepine and gnetifer trees) for each of the limited sampling data sets using substitution parameters estimated from the complete taxon sampling data sets. Finally, we explored the possibility that substitution parameters estimated from combined data sets might yield different results when used for analyses of single-locus data sets. To do this, we compared our results from ML analyses using parameters estimated from the limited taxon data sets with results from analyses of these same data sets using parameter values estimated from the combined data set.

Effect of evolutionary rate on phylogenetic signal—Previous studies of seed plant relationships have noted different phylogenetic signals in the slowly evolving first and second codon position sites when compared with the rapidly evolving third codon position sites (Chaw et al., 2000; Sanderson et al., 2000;

Magallón and Sanderson, 2002; Soltis et al., 2002). To determine the distribution of phylogenetic signal among sites evolving at different rates, we first partitioned the 13-locus combined data set using an estimate of the evolutionary rate of a particular site. We estimated the most likely rate class for each site (e.g., Yang, 1994) based on the general reversible model (Tavaré, 1986) with invariable sites and rate variation among sites following a discrete gamma distribution with eight rate categories using HYPHY (Muse and Pond, 2002). We used the MP tree to estimate the likelihood parameters including the rate classes; however, we also tried this analysis with different trees and noted very little difference in the rate class assignments for sites. This analysis partitioned the data into nine rate classes, with rate class (RC) 0 representing the invariable sites, and RC1–RC8 representing the eight discrete rate categories of the gamma distribution. RC8 represents the most rapidly evolving sites. Partitioning by rate class instead of codon position allowed us to partition the noncoding loci like 18S rDNA, 26S rDNA, and *mtSSU*, and it provided a way to standardize rate class estimates across all loci. This procedure also avoids the arbitrary placement of third codon position sites, which may evolve more slowly in some loci than in others, into the faster rate classes.

For the ML analysis, we examined the likelihood score for each site on the unconstrained tree (a gnepine tree) and on the optimal tree from a search constrained to the gnetifer topology. We calculated the likelihood difference for each site under the two hypotheses, and we examined the relationship between observed likelihood difference and the rate class of a site. For the MP analysis, we examined the parsimony score for each site on the MP tree (a Gnetales-sister tree) and on gnepine and gnetifer constraint trees. We identified the sites that have different parsimony scores for the different seed plant hypotheses, and we examined their distribution among the different estimated rate classes.

Analysis of stratigraphic data—Previous studies have noted that Gnetales as sister to all extant seed plants contradicts the stratigraphic record (Crane, 1996; Doyle, 1998a), but there are few formal tests of this observation. Doyle (1998a) found that trees with Gnetales sister to the rest had a greater percentage of ''ghost lineages'' or missing fossil data than anthophyte trees. However, Doyle (1998a) did not explicitly examine the fit of stratigraphic data to trees that join Gnetales and conifers. A likelihood-based approach allows one to calculate the significance of the likelihood ratio (δ in Fig. 8), comparing the fit of the stratigraphic data between two trees with Monte Carlo simulation (Huelsenbeck and Rannala, 1997, 2000; see also Felsenstein, 2003). We did not perform these tests because we could not estimate δ without a thorough examination of the fossil record, particularly to estimate the number of fossil observations (N). However, we did calculate likelihoods using different values of N to estimate the quantity of data that might be required to distinguish among seed plant trees using this approach. The approach is based on calculating the likelihood of observing the stratigraphic data based on a model of fossil preservation. The likelihood model assumes that fossil preservation is a Poisson process with an equal probability of fossil preservation in all lineages and across all strata. For each topology, the fossil data have the highest likelihood when the amount of evolutionary time in which there are no fossil observations is minimized. Though the model is simple, it allows information about the distribution of fossils through time to be incorporated into tests of evolutionary hypotheses. We do not have a complete data set of the stratigraphic distributions for all seed plant lineages, but we can demonstrate how the likelihood for each hypothesis varies with different numbers of fossil observations. We first mapped the times of first and last appearance of the lineages from the Doyle (1996) tree onto anthophyte, Gnetales-sister, and gnepine/gnetifer trees. The times of first appearance of angiosperms were obtained from Magallón and Sanderson (2001), and the rough estimates of first and last appearance of the seed plant lineages were obtained from Stewart and Rothwell (1993). Changing the dates of first or last observation of a few lineages may change the overall likelihood values, but in preliminary analyses they had little or no effect on the likelihood ratios (not shown). We excluded four angiosperm taxa (*Austrobaileya*, *Autunia*, *Eupomatia*, and Piperales) for which we could find no reliable appearance dates, leaving 32 lineages in each tree. Using Felsenstein's (2003, p. 554) likelihood equation, the ln likelihood of the fossil data given a tree $= -N + N(\ln N) - N(\ln T)$, where N equals the total number of fossil observations, T equals the total time length of the tree, and the average rate of fossil preservation (λ) may be estimated by N/T. The value of N is equal for all trees, but T depends on the tree topology. Because we do not know N without a complete record of the stratigraphic distribution of all lineages, we estimated the likelihoods using different N values. We chose N values ranging from 64, or two fossil observations per lineage, up to 320. The likelihood is maximized for a phylogenetic tree when T is minimized. The smallest possible T for a given tree can be calculated by finding the latest possible origin for each lineage based on the fossil record of its sister lineage (Benton and Storrs, 1994). The fit of the stratigraphic data to the gnepine and gnetifer trees is similar, and therefore, we do not examine the gnetifer tree separately.

RESULTS

Single-locus phylogenetic analyses—Trees from the single-locus data sets differ, but bootstrap support generally is low (Fig. 2). Overall, the highest bootstrap percentages support monophyly of the gymnosperms (Fig. 2). Support for a gnepine tree is highest in the *atpA* tree (100%) and lower in the *matK* (85%) and *psaA* (76%) trees (Fig. 2). Support for a gnetifer tree is highest in the 18S rDNA tree (71%; Fig. 2). The Gnetales-sister tree is not supported in ML analyses of any of the 12 loci. We also measured the likelihood ratio, or difference between the log likelihoods, between the optimal trees compatible with gnepine and gnetifer constraints ($\delta = \ln L_{GP} - \ln L_{GF}$). A positive δ indicates that the gnepine hypothesis is more likely than the gnetifer hypothesis, and a negative δ indicates the gnetifer hypothesis is more likely than the gnepine hypothesis. For eight of the 12 loci, δ was positive in all analyses, and for another, *rbcL*, δ was positive in three of the four analyses (Fig. 2). The largest δ was obtained in analyses of *atpA*, but large values were also obtained in analyses of *psaA*, and *PHYP/B* and *PHYN/A*, in each case indicating that gnepine trees were more likely than gnetifer trees. Although large values were also obtained in analyses of *mtSSU* and the 26S rDNA complete sampling data set, the ML trees were not gnepine trees (Fig. 2). For two loci, 18S rDNA and *atpB*, gnetifer trees were optimal in all analyses (Fig. 2). The ML tree from the analysis of the *rbcL* complete taxon sampling data set is also a gnetifer tree, but the gnepine tree is nearly as likely. The gnetifer hypothesis also is more likely than the gnepine hypothesis in analyses of *coxI* data, but the optimal *coxI* ML tree is neither a gnetifer nor a gnepine tree (Fig. 2).

Effect of sample size and substitution parameter values—In general, taxonomic sampling appears to have little effect on the likelihood of inferring a gnepine or gnetifer tree from single-locus data sets (Fig. 2). For 11 of the 12 loci, taxonomic sampling does not affect the sign of δ, meaning the likelihood of a gnepine relative to a gnetifer tree does not change with taxonomic sample size (Fig. 2). The exception is *rbcL* (Fig. 2). Gnepine trees inferred from the limited sampling *rbcL* data set are slightly better than gnetifer trees, whereas the reverse is true in analyses of the complete sampling data set, with $\delta < 1$ (Fig. 2). For two loci, 26S rDNA and *atpA*, differences in taxonomic sampling have a large effect on δ. The 26S data set is striking because the likelihoods of the gnepine and gnetifer trees are similar in analyses of the limited sampling data set, but the δ is by far the highest found in analyses of any complete sampling data set (Fig. 2). However, the likelihood tree inferred from the 26S complete sampling data set is not a gnepine tree and not consistent with any proposed seed plant hypothesis. The trend is reversed in analyses of *atpA*, in which

adding just seven taxa, the smallest difference between any limited and complete sampling data set pair, leads to a change in δ of 48.51 (Fig. 2).

Estimates of substitution parameters appear to have little effect on the ML analyses of the limited sampling data sets. Changing the substitution parameters in analyses of the limited sampling data sets does affect branch length estimates, and it may even affect the tree topology; however, changing the substitution parameters has little effect on the likelihood of the gnepine relative to the gnetifer hypothesis. The sign of δ never changes among results from analyses of the limited sampling data set using different parameters, and the largest change in δ due to changing parameters is 3.22 in *coxI* (Fig. 2). The percentage of invariable sites and the alpha shape parameter estimated from the complete data sets were generally lower than the estimates from the limited sampling data sets (not shown). However, the greatest change in δ due to using parameters from the complete sampling data sets in analysis of the limited taxon sampling data set is 2.71, again in analyses of *coxI* (Fig. 2).

Distribution of phylogenetic signal among rate classes—Estimation of the rate class for each site placed the majority of sites in RC0, the invariable rate class, with the mitochondrial loci having the highest proportion of invariable sites (65.1%; Table 1). There were almost no sites in RC1 or RC2 and few sites in RC3 (Table 1). The majority of variable sites were in RC4–RC8, with each rate class in this interval containing between 6.4% and 9.1% of the total sites. The rate class having the highest proportion of sites varies by genome. The nuclear loci had the most sites in the fastest rate category, RC8 (14.1%), plastid genes had the highest number in RC7 (11.1%), and the mitochondrial loci had the highest number in RC4 (10.8%; Table 1).

Although the great majority of sites in all rate classes had similar likelihoods on the gnepine and gnetifer trees, there was a greater range of likelihood differences among sites in the faster rate classes (Fig. 3). More sites in RC4 and RC5 appear to have a higher likelihood on the gnepine tree than on the gnetifer tree, but in RC6–RC8, the number of sites with a higher likelihood on the gnepine tree is similar to the number of sites that have a higher likelihood on the gnetifer tree (Fig. 3). When we summed the likelihood differences obtained for each rate class, we found that the greatest likelihood differences in favor of the gnepine tree were at RC4 and RC5 (Fig. 3). However, the highest per site average difference (δ) was in RC5 followed closely by RC3 and RC4 (Fig. 3). The variance in δ was relatively stable among rate classes, with the highest variance in rate classes 5, 6, and 7 (Fig. 3). There was virtually no overall difference between the gnepine and gnetifer hypotheses at RC8 (Fig. 3).

In the MP analysis, sites in RC1 and RC2 were uninformative with respect to the gnepine, gnetifer, and Gnetales-sister hypotheses (Fig. 4A, B). In RC3–RC6, more sites had better parsimony scores on gnepine trees than on the Gnetales-sister (Fig. 4A) or gnetifer (Fig. 4B) trees. Conversely, in RC7 and RC8, more sites had better parsimony scores on the Gnetales-sister tree than on the gnepine tree (Fig. 4A). However, when gnepine and gnetifer trees were compared (Fig. 4B), the number of sites with a better parsimony score on the gnepine tree was similar to the number of sites with a better score on the gnetifer tree. We also noted that the number of informative sites differed between the two comparisons. There were more sites that distinguish between Gnetales-sister and gnepine trees than between gnetifer and gnepine trees (Fig. 4A, B).

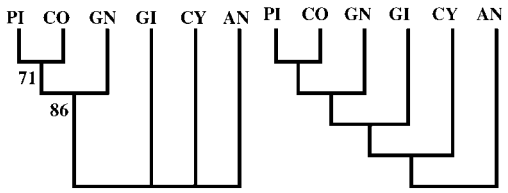
Phylogenetic analyses of combined data sets—ML analyses of all combined data sets recovered well-supported gnepine trees. The topologies of the nuclear, plastid, and mitochondrial gnepine trees (not shown) were similar to the topology of the 13-locus tree (Fig. 5). Bootstrap support for the gnepine hypothesis in trees from the different genomes ranges from 78% in the mitochondrial tree to 86% in the plastid tree and is 100% in the 13-locus tree (Fig. 5). The gymnosperms are monophyletic (98%), and the cycads are sister to the other gymnosperms (98%; Fig. 5). *Ginkgo* is sister to the Gnetales-conifer clade (100%; Fig. 5). Excluding the sites from the fastest rate classes had little effect on the ML analysis. For example, a ML analysis of the combined 13-locus data set with the RC8 sites excluded gave a tree similar to the 13-locus, also supporting the gnepine hypothesis (100%; not shown).

In contrast, results from parsimony analyses of the combined data sets were strongly affected by the choice of data partition. Data from the different genomes gave different trees, as did exclusion of the most rapidly evolving sites. The 13-locus tree (Fig. 6) and the plastid tree (Fig. 7) place Gnetales as sister to all seed plants with 79% and 100% bootstrap support, respectively. The nuclear tree places Gnetales sister to all gymnosperms with 86% bootstrap support, whereas the mitochondrial tree unites Gnetales with Pinaceae (62%; Fig. 7). However, when sites from RC7 and RC8 were excluded from MP analyses, the 13-locus tree and all single-genome trees were gnepine trees, with bootstrap support for gnepines ranging from 83% in the mitochondrial tree to 100% in the 13-locus tree (Figs. 6, 7). Cycads are sister to all other gymnosperms in the trees inferred without RC7 and RC8 (Figs. 6, 7), as they are in the ML trees (e.g., Fig. 5).

Stratigraphy—The total length of the evolutionary trees (*T*) ranged from 7564 million years in the anthophyte tree to 7795 million years in the Gnetales-sister tree. The length of the gnetifer and gnepine trees was 7595 million years. Therefore, the Gnetales-sister tree has the most fossil gaps, or time without fossil observations, and also the fossil data would have the lowest likelihood on the Gnetales-sister tree. Calculating the likelihood of the stratigraphic data requires knowing the number of fossil observations (*N*; Huelsenbeck and Rannala, 1997, 2000; Felsenstein, 2003). Lacking these data, we assumed different numbers of total fossil observations to estimate how the likelihood differences would change with various amounts of data (Fig. 8). The likelihoods of the stratigraphic data given the anthophyte and gnepine trees were close, and increasing the number of fossil observations did little to separate the likelihoods (Fig. 8). However, even assuming that there were only 64 fossil observations, or two fossils per lineage, the likelihood difference between the anthophyte or gnepine and Gnetales-sister hypothesis is nearly two. If the number of fossil observations was increased to 320, the difference in log likelihoods increased to almost 10 (Fig. 8). The 320 fossil observations (or 10 per lineage) would represent an overall fossil preservation rate (λ) of approximately one every 24 million years.

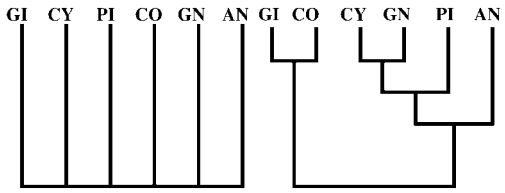
DISCUSSION

Taxonomic sample size and estimates of substitution parameters—Previous studies have suggested that increased tax-



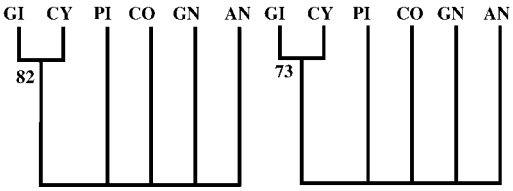
18S rDNA

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	28	28	28	96
ML	7391.77	7435.03	7503.16	12592.50
GF	7391.77	7435.03	7503.16	12592.50
GP	7396.65	7440.52	7510.16	12599.42
δ	-4.88	-5.49	-7.00	-6.93



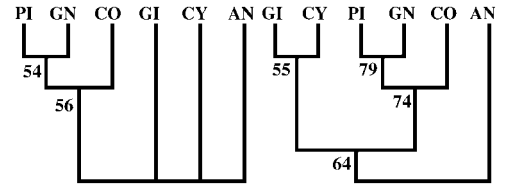
26S rDNA

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	27	27	27	98
ML	5402.79	5363.45	5443.23	8885.74
GF	5404.35	5366.33	5445.71	8944.72
GP	5402.79	5363.45	5443.23	8887.89
δ	1.57	2.88	2.48	56.83



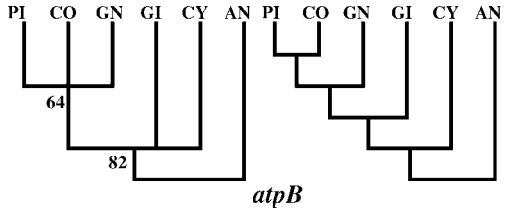
PHYP/B

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	24	24	24	40
ML	13985.78	13991.45	14024.31	22512.52
GF	13993.75	13999.66	14031.87	22525.90
GP	13985.78	13991.45	14024.31	22512.52
δ	7.97	8.21	7.56	13.37



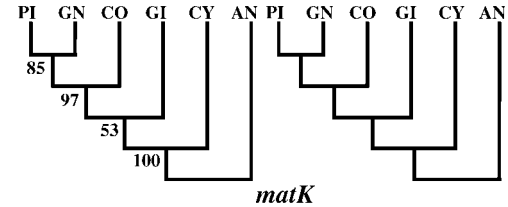
PHYN/A

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	24	24	24	43
ML	16214.38	16169.02	16260.75	24117.44
GF	16221.62	16177.16	16269.33	24126.76
GP	16214.38	16169.02	16260.75	24117.44
δ	7.24	8.14	8.59	9.31



atpB

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	29	29	29	70
ML	9727.98	9732.00	9791.83	12765.18
GF	9727.98	9732.00	9791.83	12765.18
GP	9729.07	9732.82	9793.43	12766.49
δ	-1.09	-0.82	-1.60	-1.30

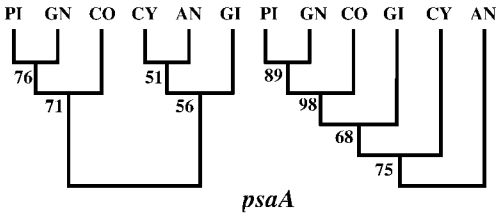


matK

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	23	23	23	240
ML	13428.18	13433.05	13920.51	26570.20
GF	13432.91	13438.01	13923.00	26577.73
GP	13428.18	13433.05	13920.51	26570.20
δ	4.73	4.96	2.49	7.53

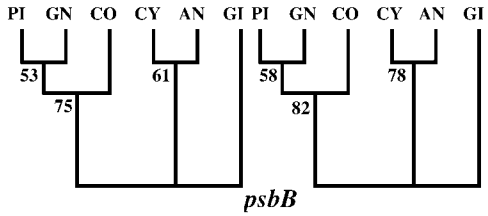
Fig. 2. The effect on seed-plant tree inference of sample size and estimates of substitution parameters for 12 loci. For each locus, trees inferred from the limited taxon data set and from the complete sampling data set are on the left and right, respectively. Maximum likelihood (ML) bootstrap percentages are on the branches of all limited taxon sampling data sets, and branches with bootstrap percentages less than 50 were collapsed. ML bootstraps are also on the complete taxon sampling trees for data sets that could be bootstrapped. In trees without bootstrap percentages, optimal trees from the likelihood searches are shown. Tables below the trees show the negative log likelihood values for unconstrained ML, the gnetifer (GF), and the gnepine (GP) constraint trees. δ represents the likelihood difference between the gnepine and gnetifer constraint trees for each analysis (= $\ln L_{GP} - \ln L_{GF}$). A positive δ means the gnepine tree has a higher likelihood than the gnetifer tree, and a negative δ means the gnetifer tree has a higher likelihood than the gnepine tree. The ''limited'' column shows the likelihood values for the limited sampling data sets with the substitution parameters estimated from the data. The ''Complete P'' column shows the likelihood scores for the limited sampling data set using substitution parameters estimated from the complete sampling data, and the ''Combined P'' column shows the likelihood score for the limited sampling data set using substitution parameters estimated from the combined 13-locus data set. The ''Complete'' column shows the likelihood scores from the complete sampling data sets. AN = angiosperms; CY = cycads; GI = *Ginkgo*; GN = Gnetales; PI = Pinaceae; CO = all other conifers. The trees are all rooted with fern or *Equisetum* outgroups.

onomic sampling may help to resolve estimates of seed plant relationships (e.g., Rydin and Källersjö, 2002; Soltis et al., 2002). We used differences in log likelihoods estimated from limited and complete taxon sampling data sets to assess whether increasing taxonomic sampling resulted in improved likelihoods, focusing on the likelihoods of gnepine and gnetifer trees. For most data sets, differences in log likelihoods between trees increased when taxa were added (Fig. 2). However, the differences were small and thus do not indicate a strong effect of taxonomic sampling on the power to choose between



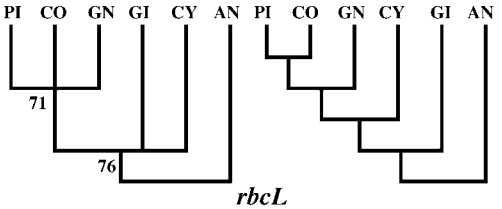
psaA

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	22	22	22	33
ML	13146.62	13147.94	13251.96	15717.79
GF	13156.15	13157.36	13262.68	15727.26
GP	13146.62	13147.94	13251.96	15717.79
δ	9.53	9.42	10.72	9.47



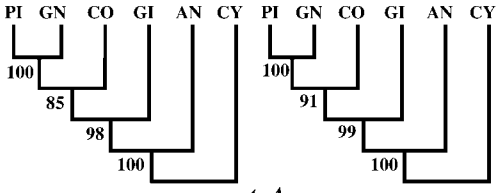
psbB

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	26	26	26	49
ML	9605.48	9638.44	9651.23	11994.54
GF	9606.27	9639.36	9652.97	11996.29
GP	9605.48	9638.44	9651.23	11994.54
δ	0.79	0.92	1.74	1.75



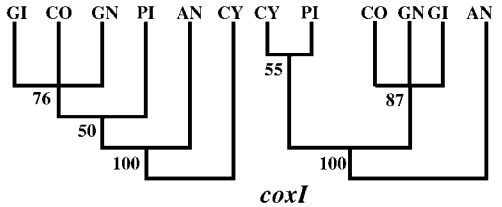
rbcL

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	29	29	29	361
ML	9049.80	9140.97	9091.67	26159.63
GF	9051.30	9141.17	9091.73	26158.43
GP	9049.80	9140.97	9091.67	26159.28
δ	1.50	0.20	0.06	-0.85



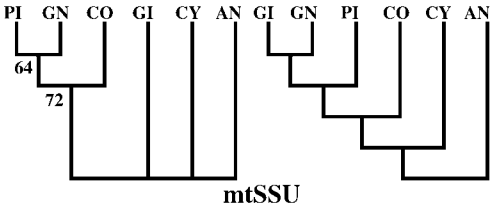
atpA

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	17	17	17	24
ML	5232.52	5233.74	5276.58	5611.32
GF	5311.92	5312.60	5357.52	5643.74
GP	5232.52	5233.74	5276.58	5611.32
δ	79.39	78.85	80.94	32.43



coxI

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	25	25	25	48
ML	6823.43	6825.46	6916.07	8368.26
GF	6838.74	6840.90	6931.50	8563.45
GP	6841.63	6841.07	6934.89	8565.17
δ	-2.89	-0.18	-3.40	-1.72



mtSSU

	Limited Sampling		Complete Sampling	
	Limited	Complete P	Combined P	Complete
Taxa	25	25	25	53
ML	6730.74	6660.06	6790.00	8112.83
GF	6746.69	6674.74	6802.75	8142.26
GP	6734.59	6663.42	6792.70	8116.43
δ	12.10	11.31	10.06	25.82

Fig. 2. Continued.

these trees. The notable departures were 26S rDNA and *mtSSU*, both of which gave unusual phylogenetic results in analyses of the complete taxon sampling data sets. Additionally, in the case of *atpA*, the smallest data set included, the difference in log likelihoods dramatically decreased when just seven taxa were added. This result appears anomalous, but it is difficult to interpret without additional sequences. Taxonomic sampling also had almost no effect on the sign of δ, or whether the gnepine or gnetifer hypothesis had a higher likelihood. In only the case of *rbcL* did the sign change from positive to negative, indicating a switch from support for a gnepine tree to support for a gnetifer tree with the addition of taxa. However, there is little overall change in δ with increased taxonomic sampling in *rbcL*, and *rbcL* by itself appears to have little power to distinguish between gnepine and gnetifer trees no matter the taxon sampling. Rydin and Källersjö (2002) noted that taxonomic sampling affected the outcomes from equally weighted parsimony analyses of an *rbcL* data set from seed plants but not from weighted parsimony or Bayesian analyses using a general time reversible model. Magallón and Sanderson (2002) also noted the limited effects of adding taxa, pointing out that key branches are missing as a result of extinctions.

Taxonomic sampling may also affect the estimates of substitution parameters (Sullivan et al., 1999). We explored the possibility that these indirect effects of taxonomic sampling on phylogenetic inference might influence the power to distinguish between gnepine and gnetifer trees and found no evidence that this was the case. Differences in log likelihoods

TABLE 1. Rate class (RC) estimates for the nuclear, plastid, and mitochondrial genomes. RC0 refers to invariable sites, and RC1–RC8 refer to the discrete rate classes in the gamma distribution, ordered from slowest to fastest. Columns list the number of sites in each rate class, with the percentage of sites per genome in parentheses.

Rate Class	Nuclear Sites	Plastid Sites	Mitochondrial Sites	Total Sites
0	3333 (55.2%)	3945 (51.7%)	3409 (65.1%)	10 687 (56.5%)
1	2 (0.0%)	2 (0.0%)	15 (0.3%)	19 (0.1%)
2	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
3	142 (2.4%)	248 (3.3%)	72 (1.4%)	462 (2.4%)
4	465 (7.7%)	674 (8.8%)	567 (10.8%)	1706 (9.0%)
5	347 (5.8%)	572 (7.5%)	292 (5.6%)	1211 (6.4%)
6	373 (6.2%)	688 (9.0%)	382 (7.3%)	1443 (7.6%)
7	523 (8.7%)	843 (11.1%)	290 (5.5%)	1656 (8.8%)
8	854 (14.1%)	657 (8.6%)	211 (4.0%)	1722 (9.1%)
Total	6039	7629	5238	18 906

among analyses of the limited sampling data sets were similar, regardless of the parameter estimates used. Together, these observations indicate that the benefits of increased sampling of extant gymnosperms may be limited. Still, our analyses do not determine if sampling the extant gymnosperms is adequate to accurately resolve seed plant phylogeny. Extinctions of major groups of seed plants have left a sparse distribution of lineages available for molecular analyses.

Conflicting phylogenetic signal—Several previous studies have revealed conflicting phylogenetic signals within single loci (e.g., Chaw et al., 2000; Sanderson et al., 2000; Magallón and Sanderson, 2002; Rydin et al., 2002; Soltis et al., 2002). To further explore these effects on the power to distinguish among trees that have been supported in multilocus analyses, we examined the phylogenetic signal in each site after partitioning all sites into nine evolutionary rate classes. Comparisons of parsimony scores on the gnepine and Gnetales-sister trees (Fig. 4a) indicate that across all loci, more sites evolving at intermediate rates favor a gnepine tree, while more sites evolving at rapid rates favor a Gnetales-sister tree. Comparisons of parsimony scores on gnepine and gnetifer trees (Fig. 4b) indicate that there are fewer sites that distinguish among these trees than in the comparison of gnepine and Gnetales-sister trees.

Comparisons of site likelihoods on the gnepine and gnetifer trees suggest that sites evolving at intermediate rates favor the gnepine hypothesis, whereas phylogenetic signal in the rapidly

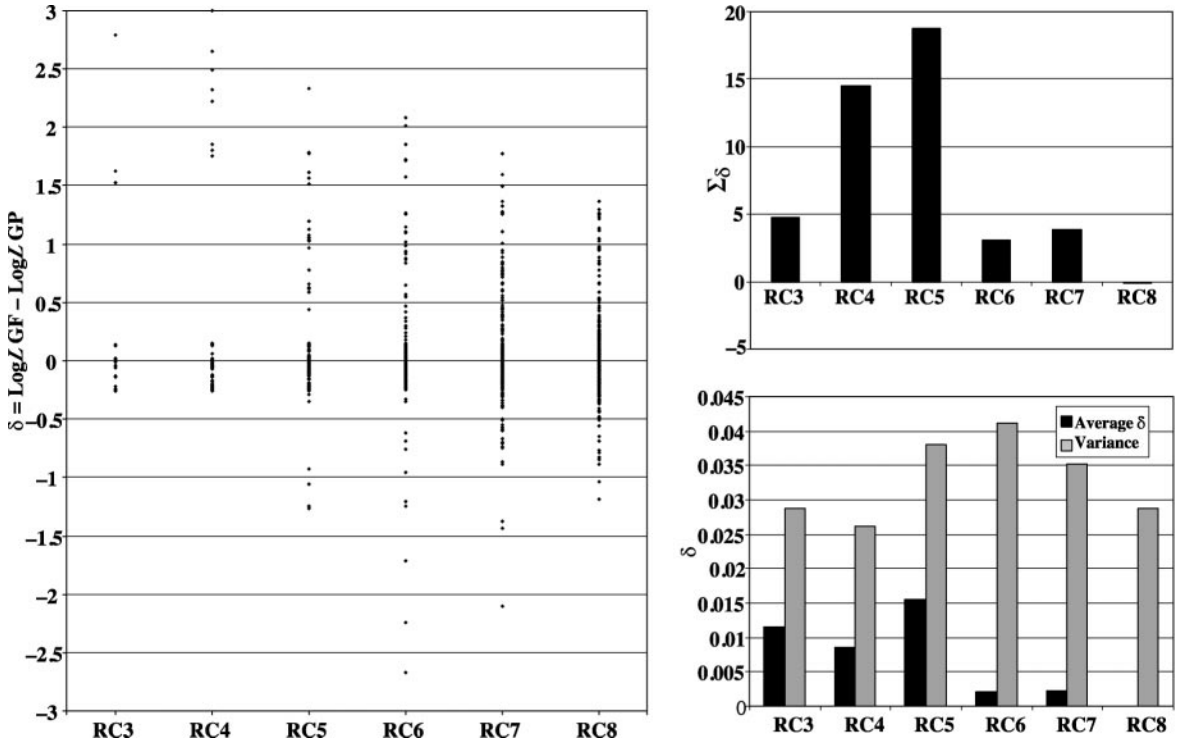


Fig. 3. Distribution of sites that have higher likelihoods for gnepine or gnetifer hypotheses among rate classes. In the graph on the left, each dot represents the ln likelihood difference at a site when optimized on the gnepine and gnetifer constraint trees inferred from the combined 13-locus data set. Positive δ values indicate that the gnepine hypothesis has a higher likelihood than the gnetifer hypothesis, and negative values indicate that the gnetifer hypothesis has a higher likelihood than the gnepine hypothesis. The smaller graph in the top right corner sums together all the δ values for the sites in rate classes 3–8, and the graph on the bottom right shows the mean and variance of the δ values for each site in rate classes 3–8.

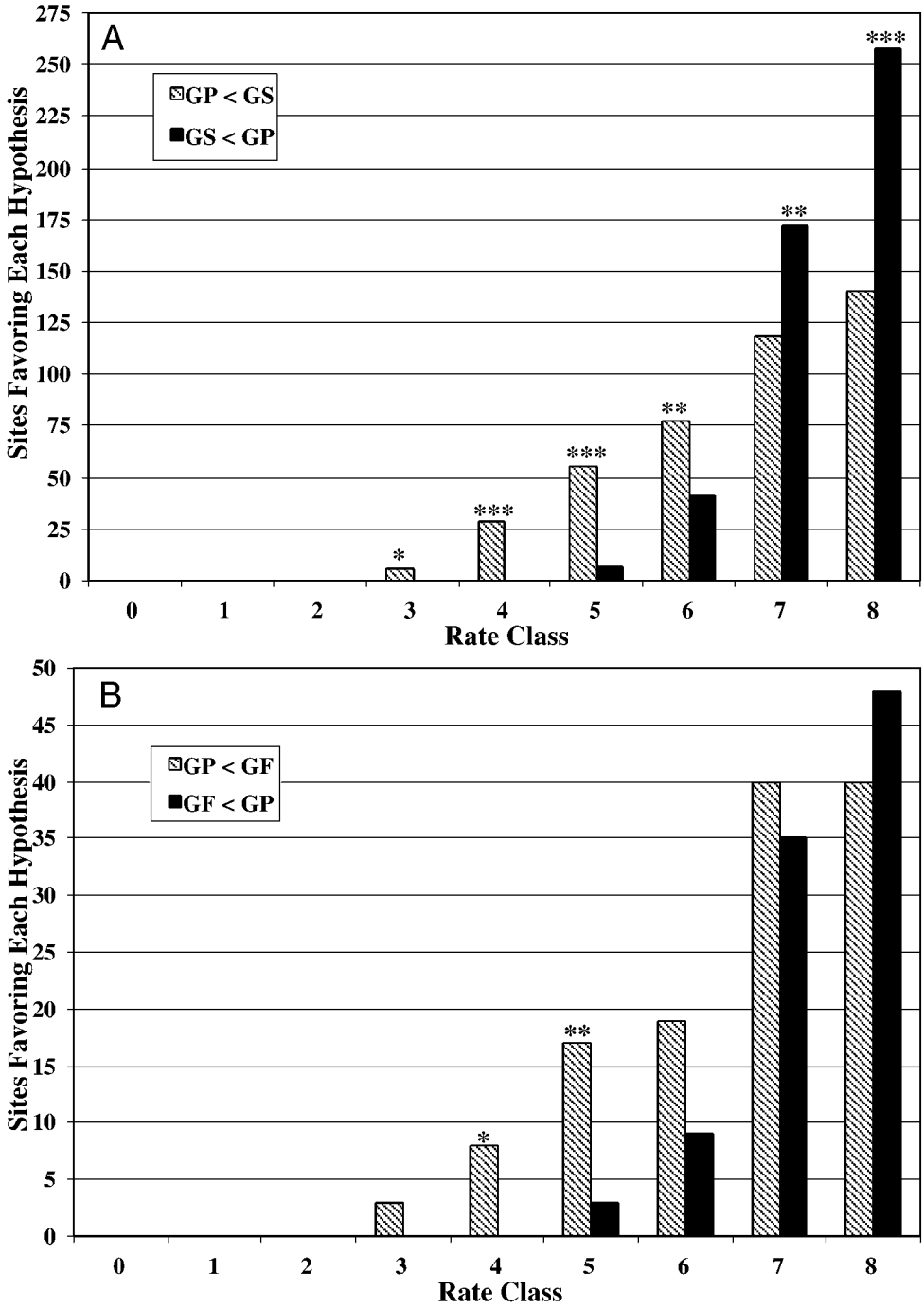


Fig. 4. Distribution of parsimony informative sites that favor gnepine (GP), gnetifer (GF), or Gnetales-sister (GS) trees among rate classes estimated using likelihood. In 4A, the striped bars indicate the number of sites in which the parsimony score for the most parsimonious gnepine tree is better than the parsimony score for the most parsimonious Gnetales-sister tree. 4B similarly compares gnepine and gnetifer trees. Chi-square tests were performed for each rate class to test the null hypothesis that the sites in which the parsimony score is variable among seed plant hypotheses are equally likely to favor either hypothesis (e.g., Snedecor and Cochran, 1995). Stars above the bars indicate that a chi-square test rejects the null hypothesis (''*'' $P \leq 0.05$, ''**'' $P \leq 0.01$, ''***'' $P \leq 0.001$).

evolving sites appears to be more ambiguous. Overall, the distribution of phylogenetic signal among rate classes is comparable to the pattern from simulated data reported by Yang (1998), who addressed the question of optimal evolutionary rate for phylogenetic inference. The slowest sites may have little phylogenetic information, whereas intermediate sites have the most, and the amount of information in the fastest sites may decrease slightly due to heterogeneity in the signal. In our analyses, the decrease in information at the most rapidly evolving sites is more pronounced than was noted by Yang (1998), and both the gnepine and gnetifer hypotheses are supported by many sites. Yang (1998) also noted that the utility

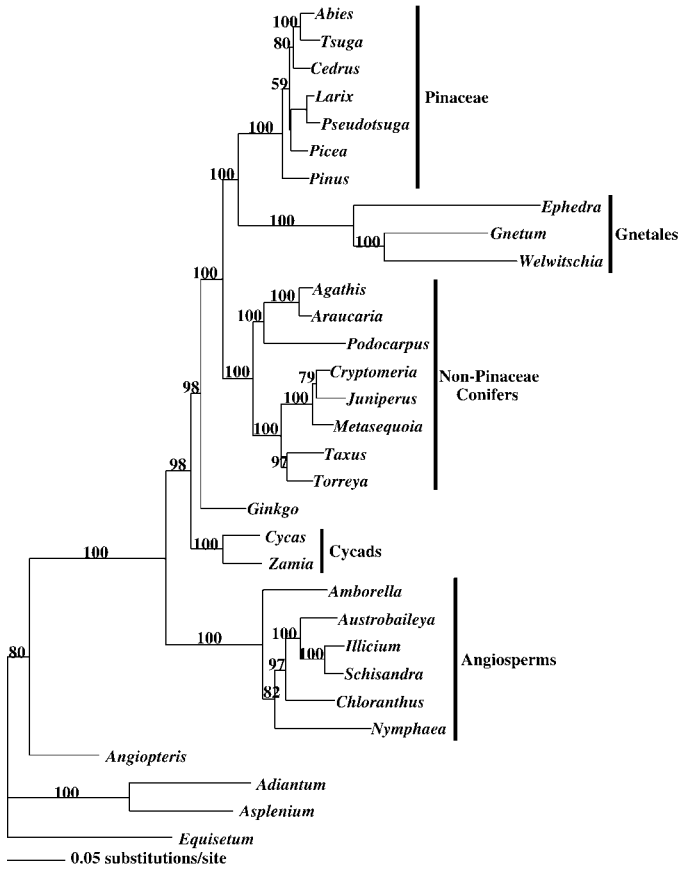


Fig. 5. Maximum likelihood tree inferred from the combined 13-locus data set. The tree is rooted with *Equisetum*. Bootstrap percentage (above 50) are placed on the branches (ln L = −124 544.1108).

of sites with different rates will vary depending on the tree shape and patterns of selection, factors that we did not explore but that are likely to be important in analyses of seed plant data. As noted for the parsimony scores, of nearly 19 000 sites in our combined data matrix, relatively few sites have greatly different likelihoods when compared on gnetifer and gnepine trees (e.g., 72 sites have $|\delta| > 1$). However, the 100% bootstrap score in the ML analysis of the 13-locus data set indicates that character sampling error is not affecting the phylogenetic inference (Fig. 5). Together, the results from exploration of the distribution of phylogenetic signal among rate classes help to explain the inconsistency in results from analyses of single-locus data sets. Considering also that Gnetales-sister trees are less consistent with stratigraphic evidence than other hypotheses (Fig. 8; Doyle, 1998a), our results indicate that the most prominent signal at the faster evolving sites is misleading in parsimony analysis.

Phylogenetic analyses of combined data—The 13-locus combined data set included a minimum of four loci from each genome, and thus is the most extensive three-genome character set so far used to address seed plant relationships. Previous analyses provided ambiguous evidence regarding the first split within gymnosperms (Bowe et al., 2000; Chaw et al., 2000; Nickrent et al., 2000; Gugerli et al., 2001; Soltis et al., 2002; Rai et al., 2003), whereas all of the deep branches in our 13-locus ML tree are supported by bootstrap percentages ≥98% (Fig. 5). Similarly, they are supported by bootstrap percentages of 100% in our 13-locus MP tree without RC7 and RC8 (Fig. 6). These trees indicate that *Ginkgo* is the sister of conifers and Gnetales and that cycads are the sister clade of all other gymnosperms (Figs. 5, 6). As in trees from other analyses, gymnosperms and angiosperms are monophyletic sister clades.

The ML tree and the MP tree from analyses without RC7 and RC8 unite Pinaceae and Gnetales with bootstrap support of 98% and 100%, respectively (Figs. 5, 6). There is no support in these trees for the suggestion that Gnetales might be embedded within Pinaceae (Soltis et al., 2002). The monophyly of Pinaceae is supported by bootstrap percentages of 100%. Previous multilocus analyses that sampled all three genomes also inferred well-supported gnepine trees in both parsimony and likelihood analyses (Qiu et al., 1999; Bowe et al., 2000; Chaw et al., 2000; Nickrent et al., 2000; Gugerli et al., 2001; Soltis et al., 2002); our 13-locus parsimony tree inferred from all sites, a Gnetales-sister tree (Fig. 6), is an exception. Multilocus analyses that sampled just the nuclear and plastid genomes (Rydin et al., 2002), or just the plastid genome (Rai et al., 2003), inferred well-supported Gnetales-sister trees in MP searches. However, our results do not support the suggestion that different trees might result from variable signal among genomes (e.g., Donoghue and Doyle, 2000; Rydin et al., 2002). All single-genome ML trees had at least 78% support for the gnepine hypothesis, and when the most rapidly evolving sites (RC7 and RC8) were excluded from parsimony analyses, all MP genome trees had at least 77% bootstrap support for the gnepine hypothesis (Fig. 7). Rather, as noted before, our results indicate that the Gnetales-sister signal is restricted to sites in the fastest two of nine evolutionary rate classes (Fig. 4A), indicating that the Gnetales-sister result may result from bias in the most rapidly evolving sites to which parsimony is particularly sensitive (Felsenstein, 1978).

Despite this convergence on the gnepine tree in molecular analyses, the result should be examined in light of evidence of possible bias or error in seed plant data sets. Sanderson et al. (2000) provided evidence of bias in first plus second codon position sites and/or in third codon position sites in two plastid data sets from seed plants and of erroneous parsimony reconstructions from these data. The bias at third codon positions was more pronounced than at first and second codon positions. The combination of very short branches at the base of the seed plant tree and very long branches leading to Gnetales and outgroups was implicated as a source of bias, and it appears that third codon positions evolve under very different processes of evolution than first and second (Sanderson et al., 2000). There may be additional factors, like model misspecification, that also could result in an erroneous phylogenetic inference. The gnepine and gnetifer topologies are very similar, and potentially even a small amount of error or bias could influence phylogenetic results. Thus, as we used stratigraphic evidence to evaluate the Gnetales-sister hypothesis, the implications of other lines of evidence should be given careful consideration.

Other sources of evidence—Despite the cladistic analyses of morphological data that supported anthophyte trees (Crane, 1985; Doyle and Donoghue, 1986, 1992; Nixon et al., 1994; Rothwell and Serbet, 1994), the clade of *Ginkgo*, Gnetales, and conifers is consistent with some morphological and anatomical evidence (Figs. 4, 5). Gnetales and conifers share a number of morphological similarities, including linear leaves, reduced sporophylls (Doyle, 1994), and circular bordered pits

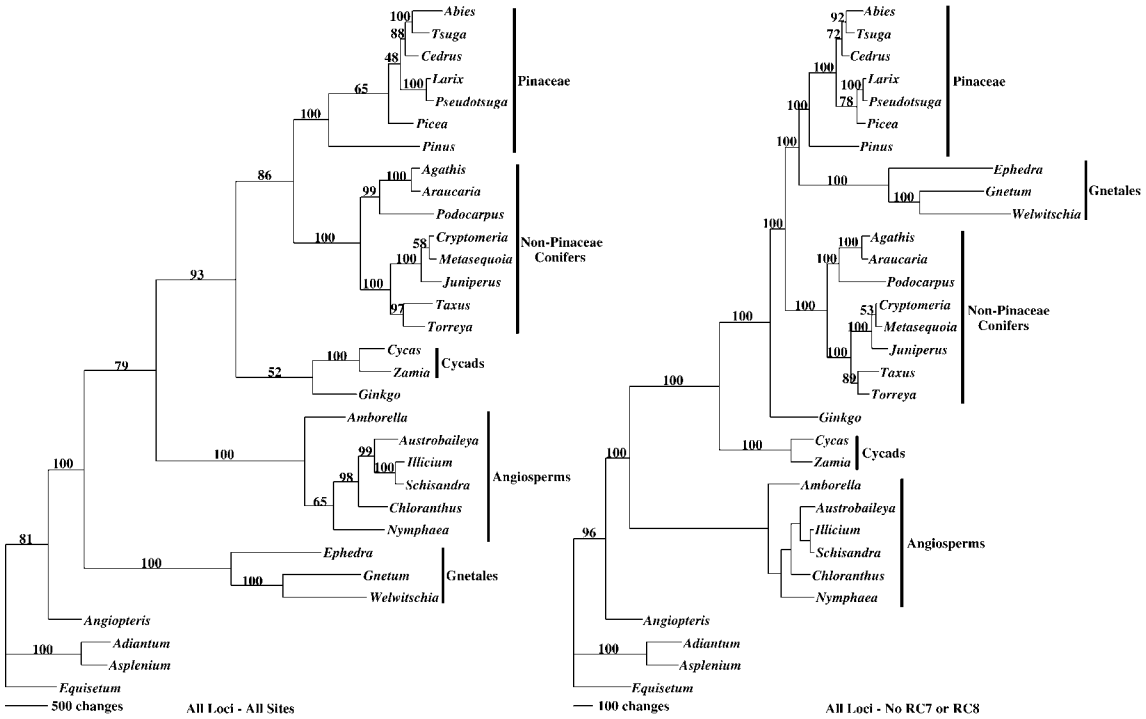


Fig. 6. Parsimony trees inferred from the combined 13-locus data set. The tree on the left is one of two equally most parsimonious trees based on an analysis that includes all sites. Bootstrap percentages above 50 are on the branches (tree length = 21 514, CI = 0.57, RI = 0.59). The tree on the right is the parsimony tree inferred from the combined data set that excludes the 3378 sites from RC7 and RC8 (tree length = 7347, CI = 0.76, RI = 0.75). All trees are rooted with *Equisetum*.

with tori in the protoxylem, interspersed with annular thickenings (Bailey, 1944; Carlquist, 1996). The similarity in the wood anatomical characters is striking and shared only with *Ginkgo* (Bailey, 1944; Bierhorst, 1971). *Ginkgo*, Gnetales, and conifers also share with extinct Cordaitales a common simple strobilar structure consisting of an axillary short shoot bearing both sterile scale leaves and simplified sporophylls. In Gnetales and Cordaitales but not *Ginkgo*, both male and female simple strobili are aggregated into compound strobili. In most conifers, female strobili are compound, whereas male strobili are simple, and this has been viewed as a feature separating conifers from Gnetales. However, in some Podocarpaceae (Wilde, 1944) and in the Paleozoic walchian conifer, *Thucydia mahoningensis* (Hernandez-Castillo et al., 2001), both male and female strobili are compound. These observations indicate that male and female compound strobili are potential synapomorphies of conifers and Gnetales. The two groups also share details of embryogeny, the presence of binucleate sperm cells, post-fertilization development of the female gametophyte, and a similar pattern of double fertilization (Friedman and Floyd, 2001). The extent to which embryogeny, fertilization patterns, and microstructural characters would support or refute the gnepine hypothesis remains to be determined in the absence of a careful survey of the distribution of the relevant character states in seed plants, and where possible, in their outgroups, but further investigation along these lines may yield data relevant to estimates of seed plant phylogeny.

The implications of gnepine trees for the evolution of other characters seem more apparent. If the gnepine tree were correct, it would suggest that a number of conifer features have evolved in parallel or were lost from Gnetales, including resin canals, tiered proembryos, and the ovulate cone scale (e.g., Chamberlain, 1935; Crane, 1985; Hart, 1987; Donoghue and Doyle, 2000). The ovules of modern conifers are borne on a woody or fleshy scale that is considered to be homologous with the lax female short shoot of Cordaitales and thus derived by a series of steps in which some appendages were lost and others congenitally fused (Wilde, 1944; Florin, 1951). If reduction of the fertile short shoot occurred once, nesting Gnetales within conifers would require the reversion of the cone scale to an axis with sporophylls subtended by bracts. The diversity of ovulate cone structure suggests that such a reversion would require a series of ontogenetic changes. Thus, it has been viewed as unlikely, and the cone scale, conversely, as a good conifer synapomorphy. However, if the female short shoot were modified independently in separate lines of conifers (Florin, 1951; Doyle, 1998b), a sister group relationship between Pinaceae and Gnetales would be less contradictory.

The causes underlying the ontogenetic diversity of ovulate cones is an important question that remains to be addressed. After an extensive survey, Tomlinson and Takaso (2002) have suggested that development is so diverse that (1) it might be taken as evidence that different types of female strobili have evolved independently or (2) that we cannot assume part-for-part homology without invoking heterochrony and heterotopy to an extent that markedly alters the model of reduction that Florin (1951) envisioned. In their view, there is strong selective pressure to protect ovules and/or seeds during development, achieved in diverse ways within seed plants, gymnosperms, and conifers themselves (Tomlinson and Takaso, 2002). If the gnepine hypothesis were correct, there seem to be two plausible alternatives. Reduction of the female short shoot may have occurred in a single lineage that is ancestral to modern conifers, followed by dissimilar elaboration of the

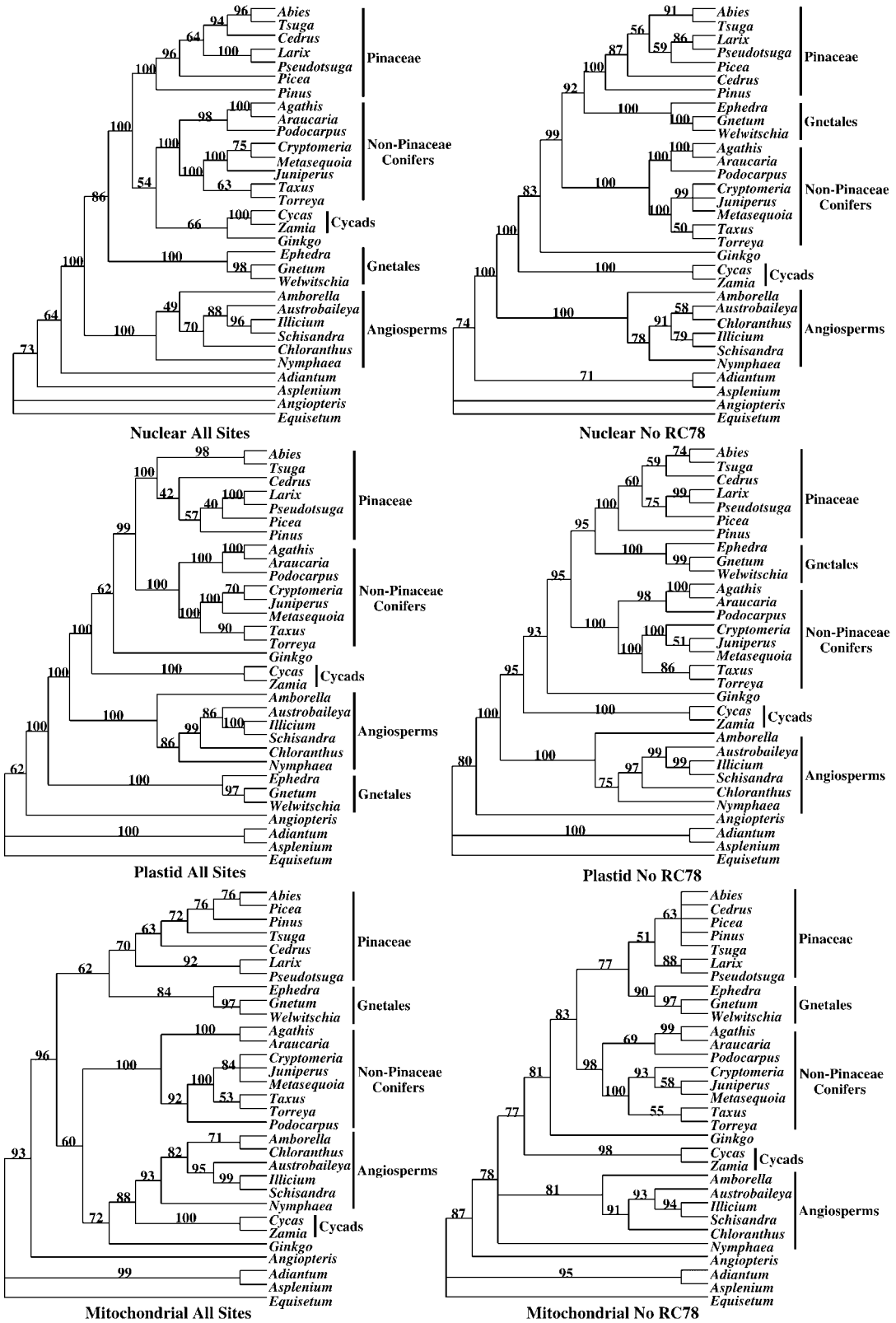


Fig. 7. Parsimony trees inferred from the combined nuclear, plastid, and mitochondrial data sets. The left column has the parsimony trees based on analyses of the nuclear (top, tree length = 8248, CI = 0.53, RI = 0.54), plastid (middle, tree length = 10 168, CI = 0.55, RI = 0.61), and mitochondrial (bottom, tree length = 2952, CI = 0.74, RI = 0.75) loci including all sites. The right column has parsimony trees based on analyses of nuclear (top, tree length = 1989, CI = 0.77, RI = 0.74), plastid (middle, tree length = 3684, CI = 0.71, RI = 0.71), and mitochondrial (bottom, tree length = 1664, CI = 0.88, RI = 0.86) loci excluding the sites from the two fastest rate classes. All trees are rooted with *Equisetum*.

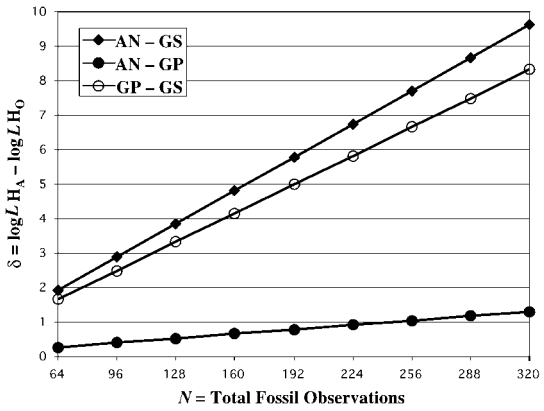


Fig. 8. The likelihood ratio test comparing the fit of stratigraphic data to seed plant hypotheses as the number of fossil observations increases. ''δ'' is the likelihood ratio of the stratigraphic data from the comparison of two different seed plant hypotheses. ''*N*'' is the total number of fossil observations across all lineages. The likelihood model assumes that fossil preservation follows a Poisson distribution with a single rate across all lineages and all strata. The ''AN—GS'' line compares the anthophyte and Gnetales-sister trees, the ''AN—GP'' lines compares the anthophyte and gnepine trees, and the ''GP—GS'' line compares the gnepine and Gnetales-sister trees.

reduced short shoot in different conifer lines, involving spatial heterogeneity in meristematic activity. Alternatively, Pinaceae and the remaining conifers may have different ancestors within Mesozoic conifers, and these may have had female short shoots that were modified independently. Molecular studies of development might provide a test of these alternatives if they could be designed to evaluate the homology of the ontogenetic pathways themselves.

Rare genomic changes (RGC; e.g., Jensen and Ahmad, 1990; Rokas and Holland, 2000) are a potential source of evidence bearing on the question of conifer monophyly and on other relationships within seed plants. Conifer monophyly apparently is supported by the absence of all or part of one copy of the plastid inverted repeat (IR) from each of the conifer families, while the Gnetales have two copies (Lidholm et al., 1988; Strauss et al., 1988; Raubeson and Jansen, 1992). A single homologous loss, rather than multiple losses, of the IR from conifers would support conifer monophyly. Sequence analysis revealed that black pine has part of a second copy, consisting of the *trnI* gene and the 3′ end of *psbA* (Tsudzuki et al., 1992; Wakasugi et al., 1994a). If this condition is widespread in conifers, it would be consistent with a homologous loss. Conversely, the absence of functional *ndh* genes from plastid genomes of both Gnetales and Pinaceae (Wakasugi et al., 1994b; Chaw et al., 2000; S. W. Graham and H. S. Rai, University of Alberta, personal communication; L. A. Raubeson, Central Washington University, personal communication) represents a potentially contradictory synapomorphy supporting gnepine trees. Other RGC appear to support the monophyly of Gnetales or clades within conifers and thus may be uninformative regarding the choice between gnepine and gnetifer trees. These include the loss of intron 2 from the mitochondrial gene *nadI* of conifers other than Pinaceae (Gugerli et al., 2001) and the absence of nuclear *PHYO* from Gnetales (Mathews and Donoghue, 2002; Schmidt and Schneider-Poetsch, 2002). The distributions of additional RGC and their utility remain to be explored, including the loss of introns from nuclear legumin genes (Shutov et al., 1998) and the invasion of IFG retrotransposons in nuclear genomes outside Pinaceae (Kossack and Kinlaw, 1999). Leads on novel, potentially useful RGC within seed plants may be obtained more efficiently as genomic data becomes more available.

Future directions—The use of sequence data to address the rooting of the seed plant tree and to identify close relatives of angiosperms will remain difficult. In analyses of molecular data, the outgroup branch consistently attaches to the seed plant tree on the branch connecting angiosperms with extant gymnosperms, suggesting that both groups are monophyletic. However, a rooting along a short internal branch of the molecular tree, such as that which separates cycads from the clade of *Ginkgo*, conifers, and Gnetales, may be particularly difficult to infer. Moreover, it is unlikely that gymnosperms as a whole are the monophyletic sister group of angiosperms as this would imply that angiosperms diverged from other seed plants as early as the Permian, well before their first appearance in the fossil record (e.g., Doyle, 1998a). Rather, it is likely that extinct taxa would attach to the branch leading to angiosperms and that gymnosperms are paraphyletic. Incomplete knowledge of the branching order and identity of the fossil taxa that diverge along this branch obscures our understanding of angiosperm origins and limits our ability to test the rooting of the seed plant tree implied by molecular trees.

Our results highlight the consistency of the gnepine result in analyses from each of the three genomic compartments and analyses that combine data from all genomes, significantly adding to the body of phylogenetic results that support the relationship between Gnetales and Pinaceae. The analyses also provide further insight into the factors leading to the inference of Gnetales-sister trees and provide criteria for discounting this result. This seems to be a clear case of erroneous parsimony reconstruction giving a tree that can be rejected in the light of nonmolecular data. However, although we have quantified a source of conflict that can be addressed to achieve a consistent result, a number of potential synapomorphies of conifers are difficult to reconcile with a gnepine tree. These bear further examination, and additional evidence should be sought from studies of morphology, anatomy, reproductive biology, and genome structure. Further exploration of the extent and source of biases and of the interactions of bias with taxon and character sampling are also needed to determine the degree to which inference of seed plant phylogeny may be influenced by biases in sequence data or analytical errors in the method of phylogenetic inference.

LITERATURE CITED

ALBERT, V. A., A. BACKLUND, K. BREMER, M. W. CHASE, J. R. MANHART, B. D. MISHLER, AND K. C. NIXON. 1994. Functional constraints and *rbcL* evidence for land plant phylogeny. *Annals of the Missouri Botanical Garden* 81: 534–567.

ANTONOV, A. S., A. V. TROITSKY, T. H. SAMIGULLIN, V. K. BOBROVA, K. M. VALIEJO-ROMAN, AND W. MARTIN. 2000. Early events in the evolution of angiosperms deduced from cp rDNA ITS 2–4 sequence comparisons. *In* Y. Liu, H. Fan, Z. Chen, Q. Wu, and Q. Zeng [eds.], Proceedings of the International Symposium on the Family Magnoliaceae, 210–214. Science Press, Beijing, China.

ARBER, E. A. N., AND J. PARKIN. 1907. On the origin of angiosperms. *Botanical Journal of the Linnean Society* 38: 29–80.

ARBER, E. A. N., AND J. PARKIN. 1908. Studies on the evolution of the angiosperms. The relationship of the angiosperms to the Gnetales. *Annals of Botany* 22: 489–515.

ARIS-BROSOU, S. 2003. Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in the presence of conflicting signals under misspecified models. *Systematic Biology* 52: 781–793.

BAILEY, I. W. 1944. The development of vessels in angiosperms and its significance in morphological research. *American Journal of Botany* 31: 421–428.

BECKER, A., AND G. THEISSEN. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution* 29: 464–489.

BENTON, M. J., AND G. W. STORRS. 1994. Testing the quality of the fossil record: paleontological knowledge is improving. *Geology* 22: 111–114.

BIERHORST, D. W. 1971. Morphology of vascular plants. Macmillan, New York, New York, USA.

BOWE, L. M., G. COAT, AND C. W. DE PAMPHILIS. 2000. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. *Proceedings of the National Academy of Science, USA* 97: 4092–4097.

BRAY, N., I. DUBCHAK, AND L. PACHTER. 2003. AVID: a global alignment program. *Genome Research* 13: 97–102.

CARLQUIST, S. 1996. Wood, bark, and stem anatomy of Gnetales: a summary. *International Journal of Plant Sciences* 157: S58–S76.

CHAMBERLAIN, C. J. 1935. Gymnosperms. Structure and evolution. University of Chicago Press, Chicago, Illinois, USA.

CHAW, S.-M., A. ZHARKIKH, H.-M. SUNG, T.-C. LAU, AND W.-H. LI. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. *Molecular Biology and Evolution* 14: 56–68.

CHAW, S.-M., C. L. PARKINSON, Y. CHENG, T. M. VINCENT, AND J. D. PALMER. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and the origin of Gnetales from conifers. *Proceedings of the National Academy of Sciences, USA* 97: 4086–4091.

CRANE, P. R. 1985. Phylogenetic analysis of seed plants and the origin of angiosperms. *Annals of the Missouri Botanical Garden* 72: 716–793.

CRANE, P. R. 1987. Vegetational consequences of angiosperm diversification. *In* E. M. Friis, W. G. Chaloner, and P. R. Crane [eds.], The origins angiosperms and their biological consequences, 107–144. Cambridge University Press, Cambridge, UK.

CRANE, P. R. 1996. The fossil history of Gnetales. *International Journal of Plant Sciences* 157: S50–S57.

CRONQUIST, A. 1968. The evolution and classification of flowering plants. Houghton Mifflin, Boston, Massachusetts, USA.

DONOGHUE, M. J., AND J. A. DOYLE. 2000. Seed plant phylogeny: demise of the anthophyte hypothesis? *Current Biology* 10: R106–R109.

DOYLE, J. A. 1978. Origin of angiosperms. *Annual Review of Ecology and Systematics* 9: 365–392.

DOYLE, J. A. 1996. Seed plant phylogeny and the relationships of Gnetales. *International Journal of Plant Sciences* 157: S3–S39.

DOYLE, J. A. 1998a. Molecules, morphology, fossils, and the relationship of angiosperms and Gnetales. *Molecular Phylogenetics and Evolution* 9: 448–462.

DOYLE, J. A. 1998b. Phylogeny of vascular plants. *Annual Review of Ecology and Systematics* 29: 567–599.

DOYLE, J. A., AND M. J. DONOGHUE. 1986. Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. *Botanical Review* 52: 321–431.

DOYLE, J. A., AND M. J. DONOGHUE. 1987. The origin of angiosperms: a cladistic approach. *In* E. M. Friis, W. G. Chaloner, and P. R. Crane [eds.], The origins angiosperms and their biological consequences, 17–49. Cambridge University Press, Cambridge, UK.

DOYLE, J. A., AND M. J. DONOGHUE. 1992. Fossils and seed plant phylogeny reanalyzed. *Brittonia* 44: 89–106.

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: S121–S153.

EAMES, A. J. 1952. Relationships of the Ephedrales. *Phytomorphology* 2: 79–100.

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. 2003. Inferring phylogenies. Sinauer, Sunderland, Massachusetts, USA.

FLORIN, R. 1951. Evolution in cordaites and conifers. *Acta Horti Bergiani* 15: 285–388.

FRIEDMAN, W. E., AND S. K. FLOYD. 2001. Perspective: the origin of flowering plants and their reproductive biology—a tale of two phylogenies. *Evolution* 55: 217–231.

FROLICH, M. W., AND D. S. PARKER. 2000. The mostly male theory of flower evolutionary origins: from genes to fossils. *Systematic Botany* 25: 155–170.

GOREMYKIN, V., V. BOBROVA, J. PAHNKE, J. TROITSKY, A. ANTONOV, AND W. MARTIN. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to the *rbcL* data do not support gnetalean affinities of angiosperms. *Molecular Biology and Evolution* 13: 383–396.

GUGERLI, F., C. SPERISEN, U. BÜCHLER, I. BRUNNER, S. BRODBECK, J. D. PALMER, AND Y.-L. QIU. 2001. The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and a multigene phylogeny. *Molecular Phylogenetics and Evolution* 21: 167–175.

HAMBY, R. K., AND E. A. ZIMMER. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. *In* P. S. Soltis, D. E. Soltis, and J. J. Doyle [eds.], Molecular systematics of plants, 50–91. Chapman and Hall, New York, New York, USA.

HANSEN, A., S. HANSMANN, T. SAMIGULLIN, A. ANTONOV, AND W. MARTIN. 1999. *Gnetum* and the angiosperms: molecular evidence that their shared morphological characters are convergent rather than homologous. *Molecular Biology and Evolution* 16: 1006–1009.

HART, J. A. 1987. A cladistic analysis of conifers: preliminary results. *Journal of the Arnold Arboretum of Harvard University* 68: 269–307.

HASEBE, M., R. KOFUKI, M. ITO, M. KATO, K. IWATSUKI, AND K. UEDA. 1992. Phylogeny of gymnosperms inferred from *rbcL* gene sequences. *Botanical Magazine, Tokyo* 105: 673–679.

HERNANDEZ-CASTILLO, G. R., G. W. ROTHWELL, AND G. MAPES. 2001. Thucydiaceae fam. nov., with a review and reevaluation of Paleozoic walchian conifers. *International Journal of Plant Sciences* 162: 1155–1185.

HUELSENBECK, J. P., AND B. RANNALA. 1997. Maximum likelihood estimation of phylogeny using stratigraphic data. *Paleobiology* 23: 174–180.

HUELSENBECK, J. P., AND B. RANNALA. 2000. Using stratigraphic information in phylogenetics. *In* J. J. Wiens [ed.], Phylogenetic analysis of morphological data, 165–191. Smithsonian Institution Press, Washington, D.C., USA.

JENSEN, R. A., AND S. AHMAD. 1990. Nested gene fusions as markers of phylogenetic branchpoints in prokaryotes. *Trends in Ecology and Evolution* 5: 219–224.

KNOLL, A. H. 1984. Patterns of extinction in the fossil record of vascular plants. *In* M. H. Nitecki [ed.], Extinctions, 21–68. University of Chicago Press, Chicago, Illinois, USA.

KOSSACK, D. S., AND C. S. KINLAW. 1999. IFG, a gypsy-like retrotransposon in *Pinus* (Pinaceae), has an extensive history in pines. *Plant Molecular Biology* 39: 417–426.

LIDHOLM, J., A. E. SZMIDT, J.-E. HÄLLGREN, AND P. GUSTAFSSON. 1988. The chloroplast genomes of conifers lack one of the rRNA-encoding inverted repeats. *Molecular and General Genetics* 212: 6–10.

LOCONTE, H., AND D. W. STEVENSON. 1990. Cladistics of the Spermatophyta. *Brittonia* 42: 197–211.

MAGALLÓN, S., AND M. J. SANDERSON. 2001. Absolute diversification rates in angiosperm clades. *Evolution* 55: 1762–1780.

MAGALLÓN, S., AND M. J. SANDERSON. 2002. Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. *American Journal of Botany* 89: 1991–2006.

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

MATHEWS, S., AND M. J. DONOGHUE. 2002. Analyses of phytochrome data from seed plants: exploration of conflicting results from parsimony and Bayesian approaches. Available at website, http://www.2002.botanyconference.org/section12/abstracts/238.shtml.

MUSE, S. V., AND S. K. POND. 2002. HYPHY: hypothesis testing using phylogenies, beta 0.95. Software program in statistical genetics, North Carolina State University, Raleigh, North Carolina, USA.

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.

NICKRENT, D. L., C. L. PARKINSON, J. D. PALMER, AND R. J. DUFF. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Molecular Biology and Evolution* 17: 1885–1895.

NIXON, K. C., W. L. CREPET, D. STEVENSON, AND E. M. FRIIS. 1994. A reevaluation of seed plant phylogeny. *Annals of the Missouri Botanical Garden* 81: 484–533.

PARENTI, L. R. 1980. A phylogenetic analysis of the land plants. *Biological Journal of the Linnean Society* 13: 225–242.

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.

POLLOCK, D. D., D. J. ZWICKL, J. A. MCGUIRE, AND D. M. HILLIS. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology* 51: 664–671.

PRYER, K. M., H. SCHNEIDER, A. R. SMITH, R. CRANFILL, P. WOLF, J. S. HUNT, AND S. D. SIPES. 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. *Nature* 409: 618–622.

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.

QUINN, C. J., R. A. PRICE, AND P. A. GADEK. 2002. Familial concepts and relationships in the conifers based on *rbcL* and *matK* sequence comparisons. *Kew Bulletin* 57: 513–531.

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.

RAUBESON, L. A., AND R. K. JANSEN. 1992. A rare chloroplast-DNA structural mutation is shared by all conifers. *Biochemical Systematics and Ecology* 20: 17–24.

ROKAS, A., AND P. W. H. HOLLAND. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution* 15: 454–459.

ROSENBERG, M. S., AND S. KUMAR. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proceedings of the National Academy of Sciences, USA* 98: 10751–10756.

ROTHWELL, G. R., AND R. SERBET. 1994. Lignophyte phylogeny and the evolution of spermatophytes: a numerical cladistic analysis. *Systematic Botany* 19: 443–482.

RYDIN, C., AND M. KÄLLERSJÖ. 2002. Taxon sampling and seed plant phylogeny. *Cladistics* 18: 484–513.

RYDIN, C., M. KÄLLERSJÖ, AND E. M. FRIIS. 2002. Seed plant relationships and the systematic position of Gnetales based on nuclear and chloroplast DNA: conflicting data, rooting problems, and the monophyly of conifers. *International Journal of Plant Sciences* 163: 197–214.

SAMIGULLIN, T. KH., W. F. MARTIN, A. V. TROITSKY, AND A. S. ANTONOV. 1999. Molecular data from the chloroplast *rpoC1* gene suggests a deep and distinct dichotomy of contemporary spermatophytes into two monophyla: gymnosperms (including Gnetales) and angiosperms. *Journal of Molecular Evolution* 49: 310–315.

SANDERSON, M. J., M. F. WOJCIECHOWSKI, J.-M. HU, T. S. 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.

SCHMIDT, M., AND H. A. W. SCHNEIDER-POETSCH. 2002. The evolution of gymnosperms redrawn by phytochrome genes: the Gnetatae appear at the base of the gymnosperms. *Journal of Molecular Evolution* 54: 715–724.

SHARROCK, R. A., AND S. MATHEWS. In press. Phytochrome genes in higher plants and their expression. *In* Photomorphogenesis in plants, 3rd ed. Kluwer Academic Publishers, Dordrect, Netherlands.

SHUTOV, A. D., H. BRAUN, Y. V. CHESNOKOV, C. HORSTMANN, I. A. KAKHOVSKAYA, AND H. BÄUMLEIN. 1998. Sequence peculiarity of Gnetalean legumin-like seed storage proteins. *Journal of Molecular Evolution* 47: 486–492.

SNEDECOR, G. W., AND W. G. COCHRAN. 1995. Statistical methods, 8th ed. Iowa State University Press, Ames, Iowa, USA.

SOLTIS, D. E., P. S. SOLTIS, AND M. J. ZANIS. 2002. Phylogeny of seed plants based on evidence from eight genes. *American Journal of Botany* 89: 1670–1681.

STEFANOVIC, S., M. JAGER, J. DEUTSCH, J. BROUTIN, AND M. MASSELOT. 1998. Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. *American Journal of Botany* 85: 688–697.

STEWART, W. N., AND G. W. ROTHWELL. 1993. Paleobotany and the evolution of plants, 2nd ed. Cambridge University Press, Cambridge, UK.

STRAUSS, S. H., J. D. PALMER, G. T. HOWE, AND A. H. DOERKSEN. 1988. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proceedings of the National Academy of Sciences, USA* 85: 3898–3902.

SULLIVAN, J., K. E. HOLSINGER, AND C. SIMON. 1996. The effect of topology on estimates of among-site rate variation. *Journal of Molecular Evolution* 42: 308–312.

SULLIVAN, J., D. L. SWOFFORD, AND G. P. NAYLOR. 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Molecular Biology and Evolution* 16: 1347–1356.

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

SWOFFORD, D. L., G. J. OLSEN, P. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. *In* D. M. Hillis, C. Moritz, and B. K. Mable [eds.], Molecular systematics, 407–425. Sinauer, Sunderland, Massachusetts, USA.

TAKHTAJAN, A. L. 1969. Flowering plants: origin and dispersal. Smithsonian Institution, Washington, D.C., USA.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *In* R. M. Miura [ed.], Lectures on mathematics in the life sciences, 5786. American Mathematics Society, Providence, Rhode Island, USA.

THORNE, R. F. 1976. A phylogenetic classification of the Angiospermae. *Evolutionary Biology* 9: 35–106.

TOMLINSON, P. B., AND T. TAKASO. 2002. Seed cone structure in conifers in relation to development and pollination: a biological approach. *Canadian Journal of Botany* 80: 1250–1273.

TSUDZUKI, J., K. NAKASHIMA, T. TSUDZUKI, J. HIRATSUKA, M. SHIBATA, T. WAKASUGI, AND M. SUGIURA. 1992. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnZ*, *trnK*, *psbA*, *trnI*, and *trnH* and the absence of *rps16*. *Molecular and General Genetics* 232: 206–214.

WAKASUGI, T., J. TSUDZUKI, S. ITO, K. NAKASHIMA, T. TSUDZUKI, AND M. SUGIURA. 1994a. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proceedings of the National Academy of Sciences, USA* 91: 9794–9798.

WAKASUGI, T., J. TSUDZUKI, S. ITO, M. SHIBATA, AND M. SUGIURA. 1994b. A physical map and clone bank of the black pine (*Pinus thunbergii*) chloroplast genome. *Plant Molecular Biology Reporter* 12: 227–241.

WETTSTEIN, R. R. 1907. Handbuch der Systematischen Botanik. Deuticke, vol. 2. Leipzig, Germany.

WIENS, J. J. 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Systematic Biology* 52: 528–538.

WILDE, M. H. 1944. A new interpretation of the coniferous cones: I. Podocarpaceae (*Podocarpus*). *Annals of Botany* 8: 1–41.

WINTER, K.-U., A. BECKER, T. MUNSTER, J. T. KIM, H. SAEDLER, AND G. THEISSEN. 1999. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. *Proceedings of the National Academy of Sciences, USA* 96: 7342–7347.

YANG, Z. 1994. Maximizing likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* 39: 306–314.

YANG, Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Systematic Biology* 47: 125–133.

YANG, Z., N. GOLDMAN, AND A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular Biology and Evolution* 11: 316–324.