Available online at www.sciencedirect.com

SCIENCE DIRECT®

Molecular Phylogenetics and Evolution 40 (2006) 208–217

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

Seed plant phylogeny: Gnetophytes are derived conifers and a sister group to Pinaceae

Mehrdad Hajibabaei [1], Junnan Xia, Guy Drouin *

Département de biologie et Centre de recherche avancée en génomique environnementale, Université d'Ottawa, Ottawa, Ont., Canada, K1N 6N5

Received 3 October 2005; revised 27 February 2006; accepted 3 March 2006
Available online 18 April 2006

Abstract

The phylogenetic position of gnetophytes has long been controversial. We sequenced parts of the genes coding for the largest subunit of nuclear RNA polymerase I, II, and III and combined these sequences with those of four chloroplast genes, two mitochondrial genes, and 18S rRNA genes to address this issue. Both maximum likelihood and maximum parsimony analyses of the sites not affected by high substitution levels strongly support a phylogeny where gymnosperms and angiosperms are monophyletic, where cycads are at the base of gymnosperm tree and are followed by ginkgos, and where gnetophytes are grouped within conifers as the sister group of pines. The evolution of several morphological and molecular characters of gnetophytes and conifers will therefore need to be reinterpreted.

Keywords: Gnetales; RNA polymerase; Molecular phylogenetics; Maximum likelihood; Maximum parsimony; Gnepines hypothesis

1. Introduction

Although many seed plant groups are known from the fossil record, only five lineages survived to the present: angiosperms, conifers, cycads, ginkgos, and gnetophytes. These last four groups are commonly known as gymnosperms because they share the putative plesiomorphic feature of a gymnospermic reproduction—with seeds that are not protected by an ovary (in contrast to angiosperms). Seed plants appeared in the late Devonian, about 370 million years ago (Kenrick and Crane, 1997). Because of the vast number of extinctions and a relatively long divergence time, inferring the evolutionary relationships of seed plants using molecular data is complicated and requires a large number of characters (Burleigh and Mathews, 2004; Magallón and Sanderson, 2002; Soltis et al., 2002; Steward and Rothwell, 1993).

The phylogenetic position of gnetophytes is one of the most controversial issues in seed plant phylogeny (Donoghue and Doyle, 2000). Morphological studies before the use of cladistics, had grouped gnetophytes with angiosperms or with conifers and *Ginkgo* (Arber and Parkin, 1908; Bailey, 1953; Wettstein, 1907). However, based on cladistic morphological analyses, gnetophytes were proposed to be the sister group of angiosperms. This hypothesis, known as the Anthophyte hypothesis (Doyle, 1996; Doyle and Donoghue, 1986), has been controversial and the focus of much research (e.g., Bowe et al., 2000; Burleigh and Mathews, 2004; Chaw et al., 2000; Donoghue and Doyle, 2000; Doyle, 1996; Goremykin et al., 1996; Graham and Olmstead, 2000; Gugerli et al., 2001; Hansen et al., 1999; Magallón and Sanderson, 2002; Rydin et al., 2002; Winter et al., 1999; Won and Renner, 2003). Overall, as recently reviewed by Burleigh and Mathews (2004), most molecular studies do not find strong support for the Anthophyte hypothesis but also do not agree on a common hypothesis regarding the relationships of gnetophytes with other gymnosperms.

Recent molecular studies have investigated the possibility of several alternative phylogenetic positions for the gnetophytes. Gnetophytes could be the sister group to all other

* Corresponding author. Fax: +1 613 562 5486.
E-mail addresses: mhajibab@uoguelph.ca (M. Hajibabaei), gdrouin@science.uottawa.ca (G. Drouin).

[1] Present address: Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, Ont., Canada, N1G 2W1.

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.03.006

seed plants, the sister group to other gymnosperms, the sister group to conifers or the sister group to Pinaceae (Bowe et al., 2000; Chaw et al., 2000; Crepet, 2000; Goremykin et al., 1996; Magallón and Sanderson, 2002; Rydin et al., 2002; Sanderson et al., 2000; Schmidt and Schneider-Poetsch, 2002; Soltis et al., 2002). This last hypothesis, known as the Gnepines hypothesis, places gnetophytes within conifers as a sister group to Pinaceae (Bowe et al., 2000; Chaw et al., 2000)—a family of conifers that forms a separate sister clade to all other conifers (Stefanovic et al., 1998). The Gnepines hypothesis has been controversial because it implies that several conifer characteristics have evolved in parallel or were lost from gnetophytes and it therefore leads to major reinterpretation of the evolution of conifers and gnetophytes (Bowe et al., 2000; Burleigh and Mathews, 2004; Chaw et al., 2000; Magallón and Sanderson, 2002).

Using sequence information to address the phylogenetic position of gnetophytes is complicated by the fact that all genes so far sequenced from this group have a relatively elevated rate of substitutions compared to other seed plant groups. This is reflected in long branches leading to gnetophytes in seed plant molecular phylogenies (e.g., Bowe et al., 2000; Chaw et al., 2000). This makes phylogenetic inference difficult because phylogenetic reconstruction methods can be subject to the phenomena of long-branch attraction where fast evolving sequences tend to be grouped together (Felsenstein, 1978). While maximum parsimony (MP) methods are particularly sensitive to this problem (Felsenstein, 1978; Hendy and Penny, 1989; Huelsenbeck, 1995), other methods, such as maximum likelihood (ML), although less influenced by long-branch attraction, are not immune to it (Huelsenbeck, 1995; Kolaczkowski and Thornton, 2004; Kuhner and Felsenstein, 1994; Swofford et al., 1996). One way to minimize this long-branch attraction problem is to break up long branches (Hillis, 1998; Zwickl and Hillis, 2002). Therefore, sampling more taxa has been suggested as a potential approach for improving seed plant phylogenetic inference (Rydin et al., 2002; Soltis et al., 2002). In the case of gnetophytes, they only contain three genera (*Ephedra*, *Gnetum*, and *Welwitschia*) and potential intermediate seed plant groups are all extinct. Therefore, the long branch leading to the gnetophytes cannot be broken by adding more extant taxa (Burleigh and Mathews, 2004). In addition, recent studies on seed plant phylogeny have shown the presence of conflicting signals, rate heterogeneity, and the sensitivity of phylogenetic methods to the choice of models of evolution and their parameters (Aris-Brosou, 2003; Bowe et al., 2000; Burleigh and Mathews, 2004; Chaw et al., 2000; Magallón and Sanderson, 2002; Rydin et al., 2002; Soltis et al., 2002). Burleigh and Mathews (2004) have recently reviewed these issues and conducted a thorough analysis of conflicting signals in available data sets. In this paper, we focused on an alternative strategy for improving phylogenetic inference of seed plant relationships, i.e., we added additional sequence data by sequencing more independent loci. In addition, we minimized the long-branch attraction and conflicting signals by identifying and analyzing relatively slowly evolving sites (Rokas and Carroll, 2005; Rosenberg and Kumar, 2001; Soltis et al., 1998).

Although a few nuclear protein-coding genes, such as homeotic and phytochrome genes, have been used to study seed plant phylogeny (e.g., Frolich and Parker, 2000; Schmidt and Schneider-Poetsch, 2002; Winter et al., 1999), most of the genes that have been used in addressing seed plant phylogeny are from organellar genomes or nuclear ribosomal genes. We generated a new data set from the genes coding for the largest subunit of RNA polymerases I, II, and III (*rpa1*, *rpb1*, and *rpc1*, respectively) as a new line of evidence and to address the lack of an extensive data set from nuclear protein-coding genes. RNA polymerase genes have been used in resolving phylogenies in a wide variety of eukaryotes (Hirt et al., 1999; Stiller and Hall, 1997; Stiller et al., 2001) including plants (Graham and Olmstead, 2000; Nickerson and Drouin, 2004). While most of these studies have used the largest and the second largest subunits of RNA polymerase II, here we obtained sequence information from the largest subunit of all three RNA polymerases. Since these three genes evolve with different rates, their combination would potentially resolve different divergence levels. In addition, genes of multi subunit protein complexes, such as RNA polymerases, may be less susceptible to phenomena such as horizontal (lateral) gene transfer (Iyer et al., 2004; Jain et al., 1999) that can complicate phylogenetic inference (Won and Renner, 2003). These genes are also often found to be single copy because the presence of numerous paralogues can cause an imbalance in the concentration of different subunits, which can be deleterious (Papp et al., 2003). Furthermore, the sequences of these three genes were obtained using expressed copies of the genes to avoid sequencing pseudogenes. This new data set was used separately and in combination with seven other available genes to study the phylogenetic position of gnetophytes.

2. Materials and methods

2.1. Data sets

This study used nine protein-coding genes, including three nuclear (*rpa1*, *rpb1*, and *rpc1*), four plastid (*psaA*, *psbB*, *rbcL*, and *atpB*), and two mitochondrial (*cox1* and *atpA*) genes, and nuclear 18S rRNA genes (total sequence length of 16,213 nucleotides) from all major groups of seed plants (12 taxa). All mitochondrial sequences were from cDNA copies used as original copies deposited in the GenBank. We avoided using a mixture of DNA and cDNA in mitochondrial genes to minimize the possible effect of RNA editing (Bowe and dePamphilis, 1996). We used *Psilotum* as the outgroup for all analyses except for mitochondrial genes, where other ferns were used as replacement taxa (see Supplementary Material). *Psilotum* was selected as outgroup because it belongs to monilifarms which has been shown to be the closest living sister group to seed plants

(Pryer et al., 2001). The sequences of chloroplast, mitochondrial, and nuclear 18S rRNA genes and seven *rpb1* genes were retrieved from GenBank. Plants species, GenBank accession numbers and references, as well as taxon substitutions to allow combined multigene analyses, are listed in the Supplementary Material.

2.2. Amplification and sequencing of RNA polymerases

To ensure sequencing functional copies of nuclear RNA polymerase genes and to avoid sequencing fast evolving intron sequences, total RNA was extracted for all 12 plant taxa using either Qiagen's RNeasy or a protocol based on the method of Bahloul and Burkard (1993). The largest subunit of RNA polymerases I, II, and III contain eight evolutionarily conserved blocks (labelled as regions A to H from 5′ to 3′; see Cramer et al. (2001) for a detailed structure). A fragment of ~3 kb between regions A and H of the *rpb1* gene of five taxa (*Thuja occidentalis*, *Ephedra viridis*, *Ginkgo biloba*, *Podocarpus macrophyllus*, and *Taxus canadensis*) was cloned and sequenced as previously described (Nickerson and Drouin, 2004). The other seven *rpb1* sequences were retrieved from GenBank. A fragment of about ~1.8 kb between regions D and G of the *rpa1* and *rpc1* genes was amplified in a single PCR using degenerate primers. A universal primer for region D of the largest subunit of all RNA polymerases was used as forward primer (Nickerson and Drouin, 2004). The sequence of this primer is: 5′-CCI TAY AAY GCI GAY ITY GAY GGI GAY GAR ATG AA-3′. However, we used a new forward primer, designed from conserved region D, for the *rpa1* gene of *Pinus nigra*: 5′-GAT GAR ATW KCW CGW GCH GAR GCH TAY AAY AT-3′. The reverse primer was designed from conserved region G of the aligned sequences of *rpa1* and *rpc1* of available taxa (mostly yeast species because plant sequences were not available at the time these primers were designed). The sequence of this primer is: 5′-GCA AAA TGA AAA GTT TTC AGA GTC ATY TGN GT-3′. The primers were designed using the CODEHOP software (Rose et al., 2003) and manual inspection.

Single-stranded cDNA was synthesized using Boehringer Mannheim's 1st Strand cDNA Synthesis Kit following the instructions of the manufacturer. RNA (0.5–0.8 μg) was used in each cDNA synthesis reaction for all plant species. Gradient PCR (with annealing temperatures of 45, 50, and 55 °C) was performed using 3 mM $MgCl_2$, 2.5 mM dNTPs, and 2 U of Taq polymerase for each reaction. PCR was repeated for 35 cycles of denaturing (94 °C for 1 min), annealing (45, 50 or 55 °C for 1 min), and extension (72 °C for 1 min). A final extension of 10 min at 72 °C was also performed at the end of the 35 cycles. PCR products were separated on an agarose gel. In the majority of cases two bands (*rpa1* and *rpc1* genes) were visible. These bands were excised from the gel (individually, when possible), purified using UltraClean15 kit (MOBIO Laboratories, Carlsbad, CA) and used in cloning. We used Invitrogen's TOPO TA cloning kit to clone each gene. Multiple clones of each gene were sequenced in both forward and reverse strands using an ABI 310 genetic analyzer. Sequences were edited and assembled using the Sequencher program (Gene Codes Corporation, Ann Arbor, MI).

2.3. Substitution and phylogenetic analyses

Sequences were aligned using ClustalW (Thompson et al., 1994) and manual inspection. The DNA sequences of protein-coding genes were aligned based on their amino acid sequence alignment. Three data sets corresponding to all three codon positions, first and second codon positions (hereafter referred to as 1 + 2), and third codon position (hereafter referred to as 3rd) were built for each protein-coding gene. The alignment of 18S rRNA genes were done using ClustalW. Regions containing numerous insertions/deletions in all sequences were not included in the analysis and gaps in single taxa were coded as missing data. The alignments of RNA polymerase genes, a combined data set of all positions of all genes (16,223 nucleotides), and a combined data set of the 1 + 2 codon positions of the RNA polymerase genes, 1 + 2 codon positions of the chloroplast genes, and all positions of mitochondrial and 18S rRNA genes (12,190 nucleotides) are available in the Supplementary Material.

The method of Pamilo–Bianchi–Li (Pamilo and Bianchi, 1993) as implemented in MEGA2.1 (Kumar et al., 2001) was used for calculating the number of synonymous and non-synonymous substitutions. We used MODELTEST (Posada and Crandall, 1998) to estimate the best evolutionary model and its parameters for each data set. These models and parameters were used in PAUP* version 4.01b (Swofford, 2002) to calculate, using a maximum likelihood approach, the number of pairwise (between individual genes, for example *rpb1* of *Ginkgo* and *rpb1* of *Psilotum*) nucleotide substitutions for 1 + 2 and 3rd codon positions.

Maximum likelihood and MP phylogenetic trees were reconstructed using PAUP* version 4.01b (Swofford, 2002) with 10 and 100 random addition TBR branch swapping tree search strategy, respectively. MODELTEST (Posada and Crandall, 1998) was used to obtain the model and parameters for the likelihood analysis for each data set. In all data sets the general time reversible model and invariable sites with gamma distribution (GTR + I + G) was selected by MODELTEST, using the AIC criterion, as the best model for the ML analyses. Statistical support for each topology was obtained using full heuristic bootstrapping (100 replicates for ML and 1000 replicates for MP with the above mentioned search strategy except that only one random addition was used in each replicate) as implemented in PAUP* version 4.01b (Swofford, 2002).

2.4. Likelihood tests of tree topologies

We used the non-parametric Shimodaira–Hasegawa (SH) test (Shimodaira and Hasegawa, 1999), as implemented in PAUP* version 4.01b, for maximum likelihood comparisons of competing tree topologies. In this test, we used the re-estimated log likelihoods (RELL) approximation with 1000 non-parametric bootstrap replicates.

2.5. Compatibility analysis

Topology independent compatibility analysis (Meacham and Estabrook, 1985; Pisani, 2004; Wilkinson, 2001) is based on the idea of Le Quesne (Le Quesne, 1969) that two characters are compatible if they can be mapped on the same tree without homoplasy. It can be used to identify fast evolving, incompatible sites within genes, and their partitions (Pisani, 2004). The test statistic, called the Le Quesne Probability (LQP), is the probability of a random character having as low or lower incompatibility with the rest of the data than does the original character (Wilkinson, 2001). We used compatibility analysis (using LQP values) to identify fast evolving sites within genes and their partitions of 1 + 2 codon positions, and 3rd codon position. We used the program DNALQP in the software package PICA 4.0 (Wilkinson, 2001) to calculate LQP values (Wilkinson, 1992) with 100 permutations. Since the compatibility testing works only on parsimony informative sites, we used these sites within each gene and gene partition for compatibility testing (Wilkinson, pers. comm.).

3. Results

3.1. Evolutionary rates of sequence partitions

We found significant evolutionary rate differences in different data sets and their partitions. Fig. 1 shows the analysis of the evolutionary rates of two data partitions in RNA polymerase, chloroplast, and mitochondrial genes. In RNA polymerase genes, most of the pairwise distances in the 3rd codon positions are above one substitution per site (Fig. 1A). In chloroplast genes, the pairwise distances in the 3rd codon positions are between 0.1 and 1.5 substitutions per site. For mitochondrial genes, most of the

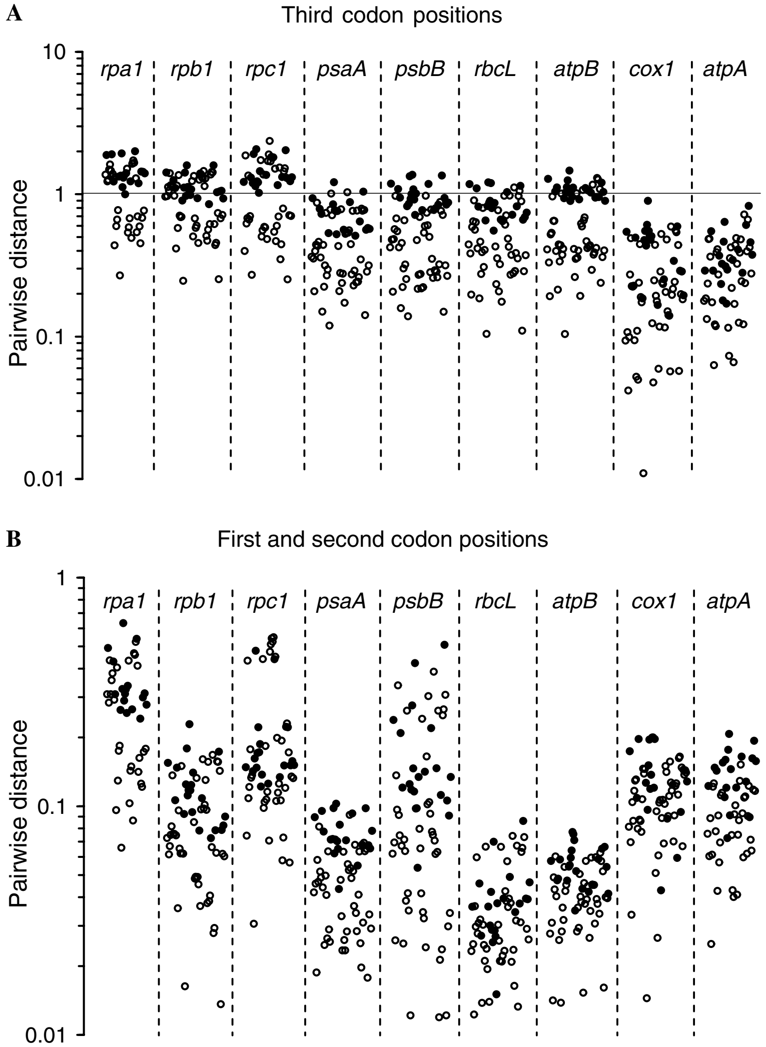

Fig. 1. Pairwise distances of 3rd (A) and 1 + 2 codon positions (B) of RNA polymerases, chloroplastic, and mitochondrial genes. Pairwise distances in which gnetophytes are one of the pairs are shown in black.

pairwise distances in the 3rd codon positions are between 0.1 and 0.9 substitutions per site. Interestingly, in RNA polymerase and chloroplast genes (but not the mitochondrial genes), most pairwise distances in which a gnetophyte gene is one of the genes are close to or above one substitution per site (Fig. 1A). Furthermore, all pairwise distances in which a gnetophyte gene is one of the genes are significantly greater than all other distance for all the genes shown in Fig. 1A (with *p*-values ranging from 0.03 for *atpA* genes 0 for *rpa1* and *cox1* genes; Table 1). In contrast, the 1 + 2 codon positions of RNA polymerases, chloroplast, and mitochondrial genes all have distance values below 0.63 substitutions per site. However, the pairwise distances in which a gnetophyte gene is one of the genes are also biased towards larger distances and are also significantly greater than all other distances for all the genes shown in Fig. 1B except for *rpc1* and *rbcL* (*p*-values range from 0.03 for *psbB* genes to 9.6×10^{-10} for *psaA* genes; Table 2).

Table 1
Comparison of the average distance of the third codon positions of gnetophyte genes with those of other taxa

Data set	Average pairwise distance of gnetophytes	Average pairwise distance of other taxa	*p*[a]
rpa1 3rd	1.442	0.991	0.000
rpb1 3rd	1.144	0.900	0.002
rpc1 3rd	1.392	1.016	0.003
psaA 3rd	0.748	0.421	2.195×10^{-7}
psbB 3rd	0.992	0.452	8.044×10^{-13}
rbcL 3rd	0.872	0.499	6.379×10^{-8}
atpB 3rd	1.067	0.549	1.971×10^{-13}
cox1 3rd	0.416	0.222	0.000
atpA 3rd	0.399	0.288	0.027
All genes 3rd	0.931	0.572	1.152×10^{-19}

[a] *p* was calculated using a two-tailed *t* test.

Table 2
Comparison of the average distance of the first and second codon positions of gnetophytes genes with those of other taxa

Data set	Average pairwise distance of gnetophytes	Average pairwise distance of other taxa	*p*[b]
rpa1 1 + 2	0.347	0.254	0.017
rpb1 1 + 2	0.120	0.080	0.001
rpc1 1 + 2[a]	0.191	0.206	0.655
psaA 1 + 2	0.076	0.043	9.630×10^{-10}
psbB 1 + 2	0.172	0.106	0.026
rbcL 1 + 2	0.039	0.033	0.197
atpB 1 + 2	0.056	0.039	2.587×10^{-06}
cox1 1 + 2	0.136	0.104	0.009
atpA 1 + 2	0.140	0.083	1.149×10^{-06}
All genes 1 + 2	0.137	0.097	2.276×10^{-05}

[a] The distances of this data set seems to be affected by the large distances involving *Nymphaea* because all the distances of *Nymphaea* and other taxa is among the higher end of the graph in Fig. 1B. Once *Nymphaea* is removed from this data set the average distances involving gnetophytes is 0.1590 which is significantly larger ($p = 0.0065$) than the average distances involving other taxa (0.125).

[b] *p* was calculated using a two-tailed *t* test.

3.2. Compatibility testing

Results from compatibility testing also suggest the presence of more conflicting sites among 3rd codon positions of RNA polymerases and chloroplast genes compared to 1 + 2 codon positions (Fig. 2). By comparing the distribution of sites with different LQP values in each gene with its partitions of 1 + 2 and 3rd codon positions, we find that, in RNA polymerases and chloroplast genes, the proportion of sites with high LQP values (i.e., more than 0.4) is higher in the 3rd codon positions compared to 1 + 2 codon positions. Therefore, removal of the third codon positions leads to a much lower number of incompatible sites in the data set compared to the data sets of all three positions (Fig. 2). In contrast, mitochondrial genes do not show such a pattern (Fig. 2).

3.3. Phylogenies inferred from different sequences and partitions

Several different phylogenies were obtained when using different genes and/or different sequence partitions. With RNA polymerase genes, the ML phylogenies based on all codon positions and on 1st and 2nd codon positions weakly support a Gnepines clade whereas the ML phylogeny based on third codon positions and the MP phylogenies do not support this clade (Fig. 3). There is therefore disagreement both between phylogenetic inference methods (ML versus MP) and partitions (1 + 2 versus 3rd codon positions). Also note that our RNA polymerase genes alignments do not contain any informative motifs (e.g., indels) which would support specific phylogenetic groupings (Hajibabaei, 2003). With chloroplast genes, the ML trees based on all codon positions and 1 + 2 codon positions and the MP tree based on 1 + 2 codon positions support the Gnepines hypothesis whereas the other three trees do not (Figure S1, Supplementary Material). With mitochondrial genes, phylogenetic analyses produced identical ML and MP trees where gymnosperms are paraphyletic (Figures S2A and B, Supplementary Material). Although these trees show the gnetophytes as being the sister group of a clade composed of *Podocarpus*, *Taxus*, and *Sequoia*, they also show *Pinus* as being the sister group of cycads and this clade as a sister group to angiosperms. ML and MP trees based on 18S rRNA genes have the same topology and show gnetophytes as being the sister group to conifers (Figures S2C and D, Supplementary Material).

3.4. Phylogenies inferred from concatenated data sets

The ML and MP phylogenetic analyses based on full sequences (total evidence) gave two different trees (Figs. 4A and B). The ML tree supports the gnetophytes as the sister group of Pinaceae (bootstrap value of 89%) whereas the MP tree supports the gnetophytes as the sister group of all other gymnosperms (bootstrap value of 93%). Since our evolutionary rate analysis showed that the third codon

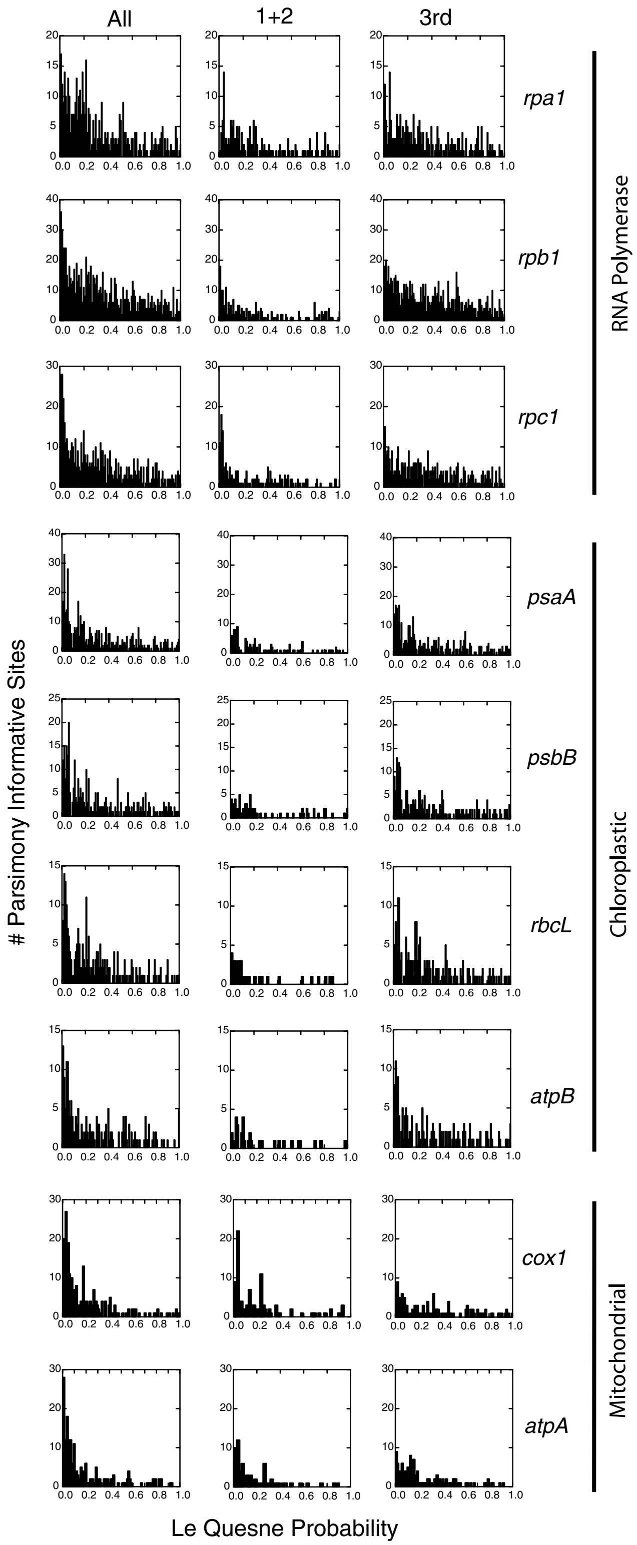

Fig. 2. Compatibility testing of each protein-coding gene and their partitions of 1 + 2 and 3rd codon positions.

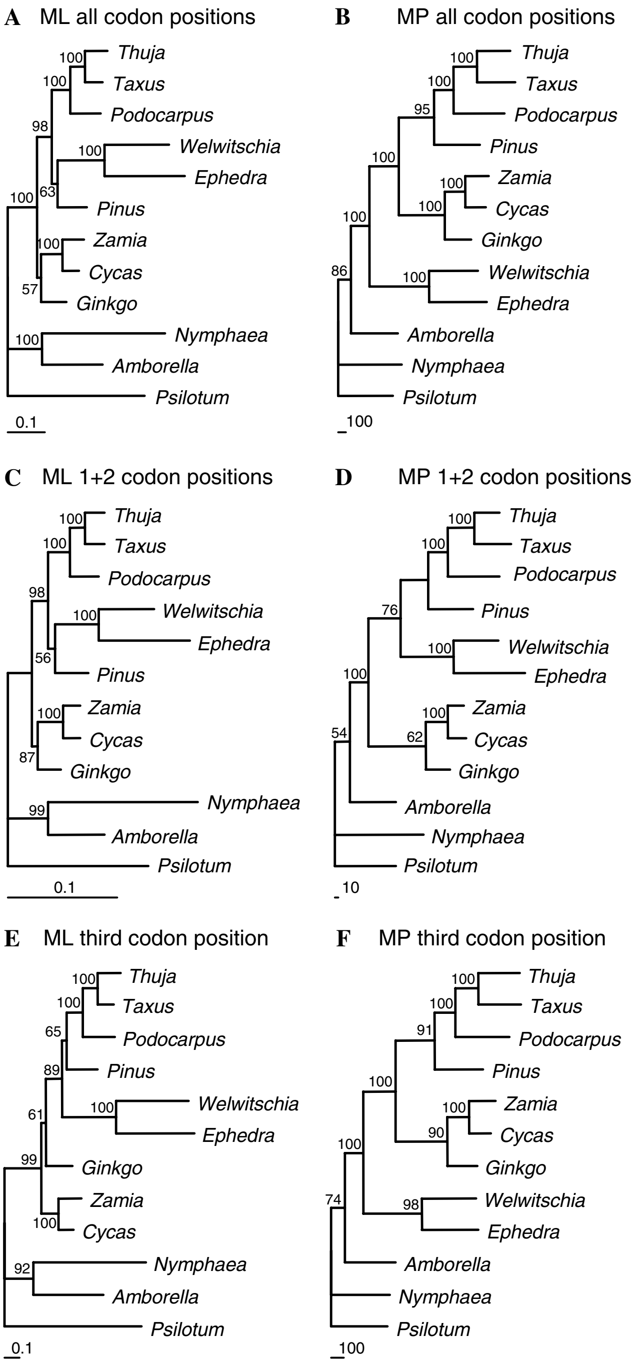

Fig. 3. Maximum likelihood (ML) and maximum parsimony (MP) phylogenies inferred from RNA polymerase genes (6273 sites) and their partitions into 1st + 2nd codon positions and 3rd codon positions. Only bootstrap values greater than 50% are shown. Scale bars represent either 10% substitution (ML) or 100 substitutions (MP).

positions of RNA polymerase and chloroplast genes have experienced high levels of substitutions (Fig. 1A), and that compatibility tests showed that the third codon positions contain relatively large number of incompatible sites (Fig. 2), we removed these sites from the analysis and combined the 1 + 2 codon positions of the RNA polymerase genes, 1 + 2 codon positions of the chloroplast genes, and all positions of mitochondrial and 18S rRNA genes (12,190 nucleotides). This data set is therefore mostly devoid of high levels of substitutions for the taxonomic level we are interested in, i.e., the phylogenetic positions of major seed plant groups. The ML and MP phylogenetic analyses based on this data set produced identical topologies that not only

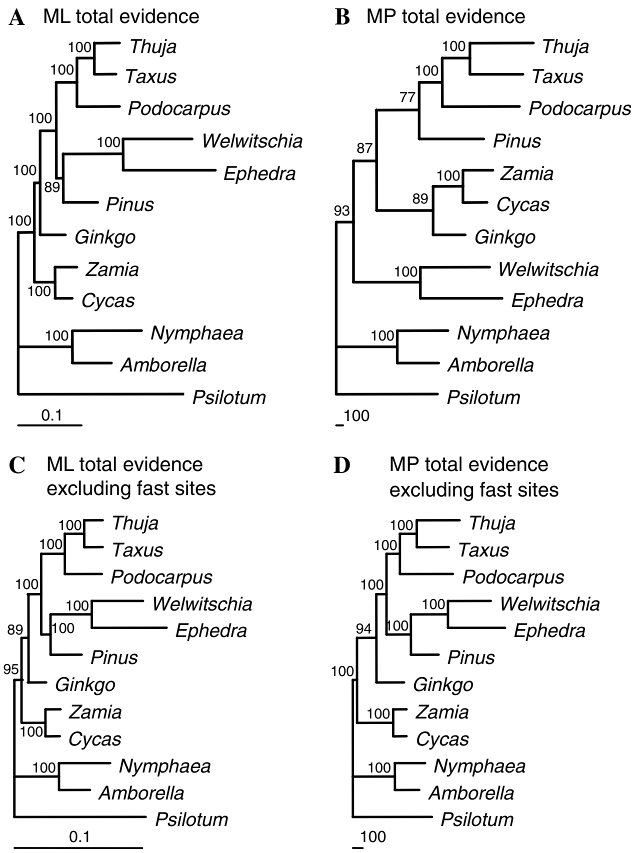

Fig. 4. Seed plant phylogenies inferred from total evidence (i.e., all positions of all genes; A and B) and total evidence excluding fast evolving sites (i.e., without third codon positions of RNA polymerase and chloroplast genes; C and D). ML, maximum likelihood; MP, maximum parsimony. Bootstrap support values are shown at the nodes and scale bars represent either 10% substitution (ML) or 100 substitutions (MP).

strongly support the Gnepines hypothesis (bootstrap values of 100%) but also strongly support that cycads are the sister group to all other gymnosperms, followed by Ginkgo and a Gnepines clade sister to the non-pine conifers (Figs. 4C and D). We also tested the effect of the outgroup on the branching order of gymnosperms in these two trees. Using angiosperms as outgroups did not change the branching pattern of the gymnosperm groups (results not shown).

3.5. Tests of tree topologies

We used the non-parametric SH-test to compare the Gnepines grouping to two other positions for gnetophytes: gnetophytes as a sister group to all other gymnosperms and, gnetophytes as a sister group to conifers (Table 3). These two alternative positions have been observed in recent phylogenetic studies (Rydin et al., 2002; Sanderson et al., 2000; Schmidt and Schneider-Poetsch, 2002; Soltis et al., 2002) as well as trees we built using different methods/data sets (e.g., Figs. 3B, D, F, and S2B). Interestingly, all SH-tests select the Gnepines grouping as the best topology (Table 3). However, the tests performed with RNA polymerase genes and chloroplast genes do not support the Gnepines grouping as being significantly better than the alternative topologies. In contrast, mitochondrial data significantly supports the Gnepines grouping. The test performed with total evidence including third codon positions of RNA polymerases and chloroplast genes indicate that the Gnepines tree is significantly better than the tree where gnetophytes are the sister group to gymnosperms but not significantly better than the tree where gnetophytes are the sister group to conifers. However, the test performed with total evidence excluding the third codon positions of RNA polymerases and chloroplast genes does support the

Table 3
Shimodaira–Hasegawa tests of alternative topologies with different data sets

Data set	Hypothesis[a]	−ln *L*	Diff. −ln *L*	*p*
RNA polymerases	[(gnetophytes) (pines)]	42181.675	Best	
	[(gnetophytes) (gymnosperms)]	42246.003	64.328	0.069
	[(gnetophytes) (conifers)]	42200.269	18.594	0.429
Chloroplast genes	[(gnetophytes) (pines)]	27196.381	Best	
	[(gnetophytes) (gymnosperms)]	27245.104	48.723	0.068
	[(gnetophytes) (conifers)]	27229.030	32.650	0.110
Mitochondrial genes	[(gnetophytes) (pines)]	11278.173	Best	
	[(gnetophytes) (gymnosperms)]	11382.904	104.731	0.000*
	[(gnetophytes) (conifers)]	11312.464	34.290	0.037*
Total evidence[b]	[(gnetophytes) (pines)]	88027.801	Best	
	[(gnetophytes) (gymnosperms)]	88245.206	217.404	0.000*
	[(gnetophytes) (conifers)]	88110.512	82.711	0.055
Total evidence without fast sites[c]	[(gnetophytes) (pines)]	45228.220	Best	
	[(gnetophytes) (gymnosperms)]	45512.790	284.570	0.000*
	[(gnetophytes) (conifers)]	45317.074	88.85410	0.011*

[a] Except for the position of gnetophytes as indicated in the hypothesis, the three topologies compared in this test are identical to the trees shown in Figs. 4C and D.

[b] All 10 genes with all codon positions.

[c] All 10 genes with all codon positions except 3rd codon positions of RNA polymerase and chloroplast genes.

* $p < 0.05$.

Gnepines grouping as being significantly better than the two alternative topologies.

4. Discussion

4.1. Selecting data partitions and the importance of more genes

Recent molecular phylogenetic studies of seed plants have produced different trees when the sequences were partitioned into 1 + 2 codon positions and 3rd codon positions or according to their substitution rates (Bowe et al., 2000; Burleigh and Mathews, 2004; Chaw et al., 2000; Magallón and Sanderson, 2002; Rydin et al., 2002; Sanderson et al., 2000; Schmidt and Schneider-Poetsch, 2002; Soltis et al., 2002). Our results confirm these observations and show that the choice of data is critical when addressing the issue of seed plant phylogeny. Using sites containing high levels of substitutions can lead the same phylogenetic method to predict incongruent trees when different data partitions are used. For example, the ML tree based on 1 + 2 codon positions of RNA polymerase genes supports the Gnepines hypothesis whereas the ML tree based on 3rd codon positions of the same genes supports gnetophytes as being the sister group of conifers (Figs. 3C and E). Using sites containing high levels of substitutions can also lead different phylogenetic methods to predict incongruent trees from the same sequence data. For example, the ML tree we obtained with all codon positions of RNA polymerase genes supports the Gnepines hypothesis whereas the MP tree we obtained with the same sequence data supports gnetophytes as being the sister group of the other gymnosperms (Figs. 3A and B). Deciding whether or not to include certain sites is not always easy because fast evolving sites have sometimes been found to be useful for building seed plants phylogenies (Burleigh and Mathews, 2004; Rydin et al., 2002; Sanderson et al., 2000; Savolainen et al., 2000; Schmidt and Schneider-Poetsch, 2002). Our results are consistent with these observations. Fig. 1 shows that although the 3rd positions of RNA polymerase and chloroplast genes are likely to contain mainly potentially misleading data at deep phylogenetic levels, the 3rd positions of mitochondrial genes are not. These results are also confirmed by an independent compatibility test (Fig. 2). Once the potentially misleading data have been removed, different phylogenetic inference methods recover the same tree (Figs. 4C and D). In addition, this tree is significantly supported by results we obtained in the SH-tests of competing topologies using a combined data set of all genes excluding the potentially misleading data partitions (Table 3). While there are a variety of tests available for comparing tree topologies, the SH-test is the most conservative (Aris-Brosou, 2003; Buckley, 2002; Goldman et al., 2000). The fact that the SH-tests support the Gnepines grouping is therefore strong support for this hypothesis. Thus, although fast evolving positions of data sets containing dense taxon sampling can be useful to resolve phylogenetic relationships, fast evolving sites have to be removed from data sets composed of more sparsely sampled taxa (Soltis et al., 2004).

We found that adding more genes improves the resolution of seed plant phylogeny and clarifies the position of gnetophytes. There are a series of studies in support of adding more taxa (Hillis, 1998; Zwickl and Hillis, 2002) or more genes (Rokas and Carroll, 2005; Rosenberg and Kumar, 2001; Soltis et al., 1998) for improving phylogenetic inference. These strategies should be employed according to the phylogenetic question at hand (Soltis et al., 2004). In the case of seed plants, the lack of lineages that break long branches due to extinctions makes it impossible to improve the phylogeny at deep levels by adding key intermediate taxa. In fact, the recent study of Burleigh and Mathews (2004) on seed plant phylogeny did not find evidence for an improved phylogeny in trees with more taxa. Our approach of sequencing three nuclear RNA polymerases genes and adding them to the available data is in line with the strategy of adding more genes to improve phylogenetic inference (Rokas and Carroll, 2005). Our results prove the effectiveness of such an approach for addressing difficult phylogenetic problems such as the position of gnetophytes.

4.2. Evolutionary relationships of seed plants

Our results suggest the following evolutionary relationships for seed plants: gymnosperms and angiosperms are monophyletic, gnetophytes are grouped within conifers as the sister group of pines, cycads are at the base of gymnosperm tree and are followed by ginkgos. The strong support we obtained for these relationships is the result of four factors. First, we added the sequence of three RNA polymerase genes which evolve relatively faster than those of chloroplast and mitochondrial genes (Table 4). They therefore allowed us to add more informative characters. Second, since *rpa1*, *rpb1*, and *rpc1* evolve at different rates from one another, and from chloroplast and mitochondrial genes (Table 4), a

Table 4
Overall mean of number of synonymous and non-synonymous substitutions per site for each protein-coding gene and data set

Gene	K_s[a]	SE[c]	K_a[b]	SE[c]
rpa1	1.063	0.045	0.189	0.012
rpb1	1.223	0.036	0.060	0.004
rpc1	1.096	0.050	0.122	0.008
psaA	0.504	0.020	0.032	0.003
psaB	0.525	0.024	0.040	0.004
rbcL	0.462	0.026	0.020	0.003
atpB	0.530	0.034	0.031	0.005
cox1	0.341	0.018	0.084	0.006
atpA	0.363	0.020	0.073	0.007
Data set				
RNA polymerases	1.188	0.025	0.097	0.003
Chloroplastic	0.502	0.012	0.031	0.002
Mitochondrial	0.341	0.014	0.081	0.005

The mean values were calculated using all the species used in this study.

[a] K_s, synonymous.

[b] K_a, non-synonymous.

[c] SE, standard error measured by bootstrapping (100 replicates).

combination of their sequences can resolve deep and shallow branches of a tree. Third, the new sequences we added were from different nuclear loci. They therefore provide not only three new data sets but these data sets are also independent from organellar and nuclear ribosomal data sets most commonly used in molecular systematic studies of plants. Finally, we identified and removed fast evolving incompatible sites (Tables 1 and 2, Fig. 2). This allowed minimizing the phylogenetic errors caused by the attraction of long branches. This proved important to resolve the phylogenetic position of the fast evolving gnetophytes.

4.3. Evolutionary implications of Gnepines grouping

Following the study of Burleigh and Mathews (2004), our study also gives 100% bootstrap support for the Gnepines hypothesis when fast evolving sites are excluded from the phylogenetic analyses (Figs. 4C and D). Furthermore, this topology is significantly better than alternative topologies (Table 3). This implies that the evolution of some molecular and morphological characters of conifers and gnetophytes will need to be reinterpreted. Although the close relationship of conifers and gnetophytes had been suggested in a number of pre-cladistic studies (Bailey, 1953; Bierhorst, 1971; Carlquist, 1996; Coulter and Chamberlain, 1917), the grouping of gnetophytes and pines nested within conifers, the Gnepines hypothesis, was only proposed using molecular data (Bowe et al., 2000; Chaw et al., 2000). Gnetophytes and conifers also possess a number of key differences such as leaf shape and growth form. Most conifers have needle shaped leaves whereas gnetophytes have a variety of leaf shapes. For example, *Gnetum* has angiosperm like leaves, *Ephedra* has scale like leaves, and *Welwitschia* has unique band-like leaves. Similarly, the growth form of conifers and gnetophytes differ widely. Gnetophytes display a notable variation of growth forms in contrast to the low diversity of growth forms found in extant conifers. Differences are also seen at the molecular level. For example, the fact that all conifers have lost one copy of the large inverted repeat of their chloroplast genome, whereas this inverted repeat is still present in gnetophytes (Raubeson and Jansen, 1992), will have to be reexamined to determine whether this pattern is due to independent losses in Pinaceae and other conifers or the reacquisition of this direct repeat by gnetophytes (Bowe et al., 2000; Chaw et al., 2000; Magallón and Sanderson, 2002). Perhaps gnetophytes once shared many morphological characters with other conifers. The accelerated morphological evolution of gnetophytes (Chaw et al., 2000; Magallón and Sanderson, 2002) correlates with the accelerated rate of evolution of their genes (Tables 1 and 2, Fig. 1). It will be interesting to determine whether this correlation reflects a causal relationship.

Acknowledgments

We thank Huguette Allard (Greenhouse, University of Ottawa), Ed Bruggink (Greenhouse, Carleton University), Brett Hall (U.C. Santa Cruz), and Lynn Gillespie (Canadian Museum of Nature) for identifying and providing plant material. We thank Harald Schneider and three anonymous reviewers for their comments on a previous version of the manuscript. This research was supported by an OGSST scholarship to M. Hajibabaei, an OGS scholarship to J. Xia and an NSERC Discovery grant to G. Drouin.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2006.03.006.

References

Arber, E., Parkin, J., 1908. Studies on the evolution of angiosperms. The relationship of the angiosperms to the Gnetales. Ann. Bot. 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. Syst. Biol. 52, 781–793.

Bahloul, M., Burkard, G., 1993. An improved method for the isolation of total RNA from spruce tissues. Plant Mol. Biol. Rep. 11, 212–215.

Bailey, I., 1953. Evolution of the tracheary tissue of land plants. Am. J. Bot. 40, 4–8.

Bierhorst, D.W., 1971. Morphology of Vascular Plants. Macmillan, New York.

Bowe, L.M., dePamphilis, C.W., 1996. Effects of RNA editing and gene processing on phylogenetic reconstruction. Mol. Biol. Evol. 13, 1159–1166.

Bowe, L.M., Coat, G., dePamphilis, C.W., 2000. Phylogeny of seed plants based on all three genomic compartments: extant gymnosperms are monophyletic and Gnetales' closest relatives are conifers. Proc. Natl. Acad. Sci. USA 97, 4092–4097.

Buckley, T.R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Syst. Biol. 51, 509–523.

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

Carlquist, S., 1996. Wood, bark, and stem anatomy of Gnetales: a summary. Int. J. Plant Sci. 157, S58–S76.

Chaw, S.M., Parkinson, C.L., Cheng, Y., Vincent, T.M., Palmer, J.D., 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc. Natl. Acad. Sci. USA 97, 4086–4091.

Coulter, J.M., Chamberlain, C.J., 1917. Morphology of Gymnosperms. The University of Chicago Press, Chicago.

Cramer, P., Bushnell, D., Kornberg, R., 2001. Structural basis of transcription: RNA polymerase II at 2.8 Å resolution. Science 292, 1863–1876.

Crepet, W.L., 2000. Progress in understanding angiosperm history, success, and relationships: Darwin's abominably "perplexing phenomenon". Proc. Natl. Acad. Sci. USA 97, 12939–12941.

Donoghue, M.J., Doyle, J.A., 2000. Seed plant phylogeny: demise of the anthophyte hypothesis. Curr. Biol. 10, R106–R109.

Doyle, J.A., 1996. Seed plant phylogeny and the relationships of Gnetales. Int. J. Plant Sci. 157, S3–S39.

Doyle, J.A., Donoghue, M.J., 1986. Seed plant phylogeny and the origin of angiosperms: an experimental cladistic approach. Bot. Rev. 52, 321–431.

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

Frolich, M.W., Parker, D.S., 2000. The mostly male theory of flower evolutionary origins: from genes to fossils. Sys. Bot. 25, 155–170.

Goldman, N., Anderson, J.P., Rodrigo, A.G., 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49, 652–670.

Goremykin, V., Bobrova, V., Pahnke, J., Troitsky, A., Antonov, A., Martin, W., 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to rbcL data do not support gnetalean affinities of angiosperms. Mol. Biol. Evol. 13, 383–396.

Graham, S.W., Olmstead, R.G., 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot. 87, 1712–1730.

Gugerli, F., Sperisen, C., Buchler, U., Brunner, I., Brodbeck, S., Palmer, J.D., Qiu, Y.L., 2001. The evolutionary split of Pinaceae from other conifers: evidence from an intron loss and a multigene phylogeny. Mol. Phylogenet. Evol. 21, 167–175.

Hajibabaei, M., 2003. Molecular evolution of the RNA polymerase genes and the phylogeny of seed plants. Ph.D. thesis. University of Ottawa.

Hansen, A., Hansmann, S., Samigullin, T., Antonov, A., Martin, W., 1999. *Gnetum* and the angiosperms: molecular evidence that their shared morphological characters are convergent, rather than homologous. Mol. Biol. Evol. 16, 1006–1009.

Hendy, M., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38, 297–309.

Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47, 3–8.

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

Huelsenbeck, J., 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44, 17–48.

Iyer, L.M., Koonin, E.V., Aravind, L., 2004. Evolution of bacterial RNA polymerase: implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer. Gene 335, 73–88.

Jain, R., Rivera, M.C., Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. Proc. Natl. Acad. Sci. USA 96, 3801–3806.

Kenrick, P., Crane, P.R., 1997. The origin and early evolution of plants on land. Nature 389, 33–39.

Kolaczkowski, B., Thornton, J.W., 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431, 980–984.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.

Kumar, S., Tamura, K., Jakobsen, I.B., Nei, M., 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245.

Le Quesne, W.J., 1969. A method of selection of characters in numerical taxonomy. Syst. Zool. 18, 201–205.

Magallón, S., Sanderson, M.J., 2002. Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. Am. J. Bot. 89, 1991–2006.

Meacham, C.A., Estabrook, G., 1985. Compatibility methods in systematics. Annu. Rev. Eco. Syst. 16, 431–446.

Nickerson, J., Drouin, G., 2004. The sequence of the largest subunit of RNA polymerase II is a useful marker for inferring seed plant phylogeny. Mol. Phylogenet. Evol. 31, 403–415.

Pamilo, P., Bianchi, N.O., 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol. Biol. Evol. 10, 271–281.

Papp, B., Pal, C., Hurst, L.D., 2003. Dosage sensitivity and the evolution of gene families in yeast. Nature 424, 194–197.

Pisani, D., 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53, 978–989.

Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14, 817–818.

Pryer, K.M., Schneider, H., Smith, A.R., Cranfill, R., Wolf, P.G., Hunt, J.S., Sipes, S.D., 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature 409, 618–622.

Raubeson, L., Jansen, R., 1992. A rare chloroplast-DNA structural mutation is shared by all conifers. Biochem. Syst. Ecol. 20, 17–24.

Rokas, A., Carroll, S.B., 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22, 1337–1344.

Rose, T.M., Henikoff, J.G., Henikoff, S., 2003. CODEHOP (COnsensus-DEgenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res. 31, 3763–3766.

Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98, 10751–10756.

Rydin, C., Källersjö, M., Friis, E.M., 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. Int. J. Plant Sci. 163, 197–214.

Sanderson, M.J., Wojciechowski, M.F., Hu, J.M., Khan, T.S., Brady, S.G., 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol. Biol. Evol. 17, 782–797.

Savolainen, V., Chase, M.W., Hoot, S.B., Morton, C.M., Soltis, D.E., Bayer, C., Fay, M.F., de Bruijn, A.Y., Sullivan, S., Qiu, Y.L., 2000. Phylogenetics of flowering plants based on combined analysis of plastid *atpB* and *rbcL* gene sequences. Syst. Biol. 49, 306–362.

Schmidt, M., Schneider-Poetsch, H.A.W., 2002. The evolution of gymnosperms redrawn by phytochrome genes: the *Gnetatae* appear at the base of the gymnosperms. J. Mol. Evol. 54, 715–724.

Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16, 1114–1116.

Soltis, D.E., Soltis, P.S., Mort, M.E., Chase, M.W., Savolainen, V., Hoot, S.B., Morton, C.M., 1998. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. Syst. Biol. 47, 32–42.

Soltis, D.E., Soltis, P.S., Zanis, M., 2002. Phylogeny of seed plants based on evidence from eight genes. Am. J. Bot. 89, 1670–1681.

Soltis, D.E., Albert, V.A., Savolainen, V., Hilu, K., Qiu, Y.L., Chase, M.W., Farris, J.S., Stefanovic, S., Rice, D.W., Palmer, J.D., Soltis, P.S., 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. 9, 477–483.

Stefanovic, S., Jager, M., Deutsch, J., Broutin, J., Masselot, M., 1998. Phylogenetic relationships of conifers inferred from partial 28S rRNA gene sequences. Am. J. Bot. 85, 688–697.

Steward, W.N., Rothwell, G.W., 1993. Paleobotany and the Evolution of Plants, second ed. Cambridge University Press, Cambridge.

Stiller, J.W., Hall, B.D., 1997. The origin of red algae: implications for plastid evolution. Proc. Natl. Acad. Sci. USA 94, 4520–4525.

Stiller, J.W., Riley, J., Hall, B.D., 2001. Are red algae plants? A critical evaluation of three key molecular data sets. J. Mol. Evol. 52, 527–539.

Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Computer program and documentation, Sinauer Associates, Sunderland MA.

Swofford, D.L., Olsen, G.L., Waddell, P.J., Hillis, D.M., 1996. Molecular Systematics, second ed. Sinauer Associates, Sunderland MA.

Thompson, J.D., Higgins, D.G., Gibson, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.

Wettstein, R.R., 1907. Handbuch der Systematischen Botanik Deuticke. Leipzig, Germany.

Wilkinson, M., 1992. Consensus, Compatibility and Missing Data in Phylogenetic Inference. University of Bristol, Bristol.

Wilkinson, M., 2001. PICA 4.0: Software and documentation. The Natural History Museum.

Winter, K.U., Becker, A., Munster, T., Kim, J.T., Saedler, H., Theissen, G., 1999. MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. Proc. Natl. Acad. Sci. USA 96, 7342–7347.

Won, H., Renner, S.S., 2003. Horizontal gene transfer from flowering plants to *Gnetum*. Proc. Natl. Acad. Sci. USA 100, 10824–10829.

Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598.