Chloroplast Phylogenomics Indicates that *Ginkgo biloba* Is Sister to Cycads

Chung-Shien Wu[1], Shu-Miaw Chaw[1,*], and Ya-Yi Huang[1,2]

[1]Biodiversity Research Center, Academia Sinica, Taipei, Taiwan

[2]Present address: Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan

*Corresponding author: E-mail: smchaw@sinica.edu.tw.

Accepted: January 4, 2013

Data deposition: GenBank accession numbers JQ770198–JQ770303; DDBJ accession numbers AB715148–AB715200 and AB715249–AB715301.

Abstract

Molecular phylogenetic studies have not yet reached a consensus on the placement of Ginkgoales, which is represented by the only living species, *Ginkgo biloba* (common name: ginkgo). At least six discrepant placements of ginkgo have been proposed. This study aimed to use the chloroplast phylogenomic approach to examine possible factors that lead to such disagreeing placements. We found the sequence types used in the analyses as the most critical factor in the conflicting placements of ginkgo. In addition, the placement of ginkgo varied in the trees inferred from nucleotide (NU) sequences, which notably depended on breadth of taxon sampling, tree-building methods, codon positions, positions of Gnetopsida (common name: gnetophytes), and including or excluding gnetophytes in data sets. In contrast, the trees inferred from amino acid (AA) sequences congruently supported the monophyly of a ginkgo and Cycadales (common name: cycads) clade, regardless of which factors were examined. Our site-stripping analysis further revealed that the high substitution saturation of NU sequences mainly derived from the third codon positions and contributed to the variable placements of ginkgo. In summary, the factors we surveyed did not affect results inferred from analyses of AA sequences. Congruent topologies in our AA trees give more confidence in supporting the ginkgo–cycad sister-group hypothesis.

Key words: phylogenomics, cycads, chloroplast, seed plants, ginkgo.

Introduction

Chloroplasts are usually uniparentally inherited and have their own genomes, with multiple genome copies per chloroplast, dense gene content, and slower evolutionary rates than those of nuclear and mitochondrial genomes (Wolfe et al. 1987; Drouin et al. 2008). Recent advances in next-generation sequencing technologies have facilitated the rapid assembly of complete chloroplast genomes from total DNA (Nock et al. 2011). This development has increased the number of complete chloroplast genomes in genetic databases. Furthermore, phylogenetic analyses based on chloroplast genomes of representative species have become a popular and practical approach (e.g., Jansen et al. 2007; Moore et al. 2007, 2010; Lin et al. 2010; Wu et al. 2011). This approach called "chloroplast phylogenomics" provides a means to test previously suggested hypotheses (Martin et al. 2005).

Nonetheless, phylogenomics has some limitations. For example, Jeffroy et al. (2006) asserted that large-scale genomic approaches might result in significant incongruence among methods and character sets. Philippe et al. (2011) claimed that adding more sequences is insufficient for resolving difficult phylogenetic questions because "non-phylogenetic" signals derived from model violation (systematic errors) might increase and compete with "genuine" signals. Indeed, recent chloroplast phylogenomic studies (e.g., Goremykin et al. 2009, 2013; Zhong et al. 2011) suggested that the accuracy of phylogenomic analyses can be improved by removal of model violating sites (e.g., the most variable sites) from data sets. However, if a true topology is unknown, a criterion is required to evaluate the accuracy of trees inferred from a set of real sequences. Delsuc et al. (2005) suggested that "the congruence of results obtained from various data sets and/or various methods is the key validation of evolutionary inferences." Leigh et al. (2011) supported this view and stated that "congruence is a broadly applied notion in evolutionary biology used to justify multigene phylogeny or phylogenomics." We propose that topological congruence

increases our confidence in inferred trees and provides a means for identifying factors that may lead to phylogenetic incongruence. Here, the previously controversial placements of ginkgo were used as a case study.

Ginkgo is the sole genus of Ginkgoales and represents a lineage that has existed for at least 270 Myr. Despite considerable studies, the relationships of ginkgo to the other four extant gymnosperm lineages—cycads, pine family (Pinaceae), cupressophytes (non-Pinaceae Coniferales), and gnetophytes—have not been resolved. For example, Norstog et al. (2004) suggested that ginkgo is closer to conifers than cycads, whereas Wang et al. (2011) concluded that ginkgo is intermediate between these two lineages. To date, molecular studies have reported six possible placements for ginkgo. These range from being sister to cycads to sister to all the remaining gymnosperms (fig. 1). Unfortunately, little attention has been paid to examine why resolving the phylogenetic placement of this living fossil has been problematic.

Figure 1 clearly depicts different breadth of taxon sampling and diverse positions of gnetophytes in previous studies. Whether these two factors are determinants in the placement of ginkgo needs to be investigated. Particularly, genes of the gnetophyte chloroplast genomes have extremely accelerated rates with high AT content (Wu et al. 2009), which is the source of long-branch attraction (LBA). In addition, different tree-building methods (maximum likelihood [ML], Bayesian inference [BI], and maximum parsimony [MP]) and sequence types (nucleotide [NU] and amino acid [AA]) might contribute to the incongruent placements of ginkgo in previous analyses.

The chloroplast genome sequence of ginkgo has recently been completed (Lin et al. 2012). This study aims to investigate factors that may have led to uncertainty in the phylogenetic placement of ginkgo and re-evaluate this long-debated issue. Therefore, our analyses were based on the largest and most comprehensive sampling of available gymnosperm chloroplast genomes (25 species), including two species of Zamiaceae (Cycadalesles), one Indian gnetophyte and one cupressophyte that are reported for the first time. We expected that this increased sampling would help us to better evaluate the phylogenetic placement of ginkgo and to explain the evolution of protein-coding genes of chloroplast genomes.

Materials and Methods

DNA Extraction

Young leaves were harvested from plants of *Ceratozamia hildae* (Cycadales, Zamiaceae), *Zamia furfuracea* (Cycadales, Zamiaceae), *Calocedrus formosana* (Coniferales, Cupressaceae), and *Gnetum ula* (Gnetales, Gnetaceae) in the greenhouse of Academia Sinica, Taiwan. Fresh leaf tissue (2 g) was ground under liquid nitrogen and used for DNA extraction. Total DNA was extracted using a 2× CATB method (Stewart and Via 1993) with 0.01 g/ml PVP-40. The extracted DNA was determined by the threshold—DNA concentration >300 ng/μl, 260/280 ratio = 1.8–2.0, and 260/230 ratio >1.7—on a GeneQuant 1300 spectrophotometer, Biochrom Ltd.

Sequencing, Sequence Assembly, and Gene Annotation

The DNA samples were sequenced by Yourgene Bioscience (New Taipei City, Taiwan) using an Illumina GAII instrument to generate 73-bp paired-end reads. For each species, the short reads were trimmed with the threshold of error probability < 0.05 and were de novo assembled using CLC Genomic Workbench 4.9 (CLC Bio, Aarhus, Denmark). Contigs <200 bp long were discarded. The remaining contigs were compared with all plant complete chloroplast genomes in the National Center for Biotechnology Information (NCBI) database using a BLAST search. Contigs that matched referenced genomes with *E* values $< 10^{-5}$ were retained for annotation. Mean coverages of the contigs were estimated to be 82×, 32×, 34×, and 201× for *C. hildae*, *Z. furfuracea*, *Cal. formosana*, and *G. ula*, respectively. Protein-coding genes were annotated using DOGMA (Wyman et al. 2004). Boundaries of each annotated gene were manually compared with orthologous genes from other gymnosperm chloroplast genomes.

Alignments and Construction of Data Sets

We extracted 53 common chloroplast protein-coding genes (supplementary table S1, Supplementary Material online) for 60 sampled taxa (supplementary table S2, Supplementary Material online) from the NCBI organelle database. With our newly elucidated sequences of the four gymnosperms, multiple sequence alignments for each orthologous gene were performed using MUSCLE (Edgar 2004) as implemented in Mega 5 (Tamura et al. 2011). Alignments of the 53 genes were concatenated manually. We used Mega 5 to convert NU to AA and excluded all gaps and ambiguous sites. The two resulting matrices included 35,994 NU and 11,998 AA positions (supplementary file S1, Supplementary Material online). To examine the influence of taxon sampling, the two full matrices (NU and AA data set 3; 56 seed plants, 7 ferns, and 1 lycophyte) were subsampled in two ways. One subsample contained only the sampled seed plants (NU and AA data set 1; 56 seed plants) and the other the sampled seed plants and ferns (NU and AA data set 2; 56 seed plants and 7 ferns).

Model-Test Analyses

For each data set, the best-fit model for NU and AA sequences were evaluated using jModelTest 2 (Darriba et al. 2012) and ProtTest 3 (Darriba et al. 2011) without data partitioning, respectively. Both Akaike Information Criterion and Bayesian Information Criterion were used to compare the examined models (88 models for NU and 120 models for AA) and find

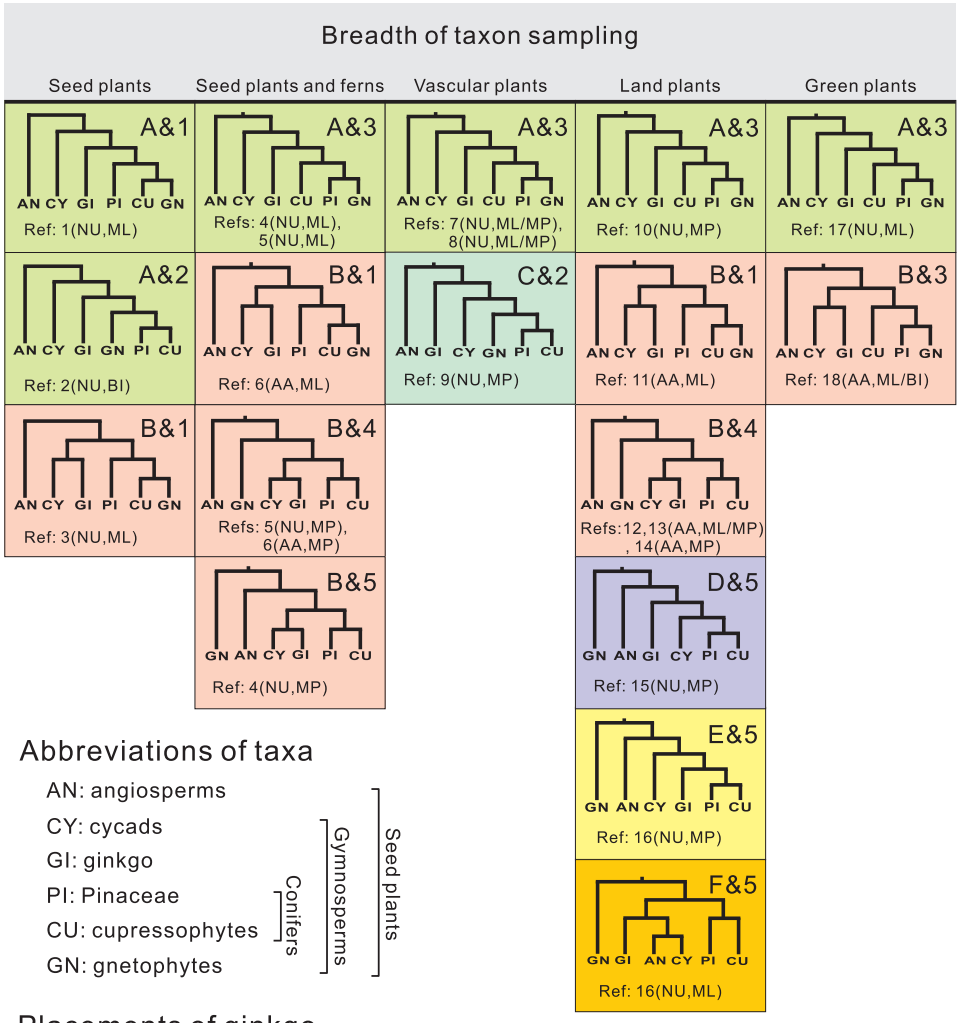

Fig. 1.—Previously reported placements of *Ginkgo biloba*. The six competing placements (A–F) are highlighted with different colored backgrounds. Indicated factors include breadth of taxon sampling, positions of gnetophytes (numbers 1–5), sequence type (NU and AA), and tree-building method (ML, BI, and MP). Reference (Ref): 1, Graham and Iles (2009); 2, Ran et al. (2010); 3, Zhong et al. (2011); 4, Burleigh and Mathews (2004); 5, Hajibabaei et al. (2006); 6, Wu et al. (2011); 7, Bowe et al. (2000); 8, Chaw et al. (2000); 9, Chaw et al. (1997); 10, Regina and Quagliariello (2010); 11, Zhong et al. (2010); 12, de la Torre-Bárcena et al. (2009); 13, Lee et al. (2011); 14, Cibrián-Jaramillo et al. (2010); 15, Rydin et al. (2002); 16, Rai et al. (2008); 17, Qiu et al. (2007); and 18, Finet et al. (2010).

the best-fit models. The results were shown in supplementary table S3, Supplementary Material online.

Phylogenomic Analyses

ML, BI, and MP searches were conducted using RAxML 7.2.8 (Stamatakis 2006), MrBayes 3.2.1 (Ronquist et al. 2012), and PAUP 4.0b (Swofford 2002), respectively. For ML analyses, the best-fit models, general time reversible (GTR) + G_4 + I and Jones, Taylor, and Thorton (JTT) + G_4 + F models (supplementary table S3, Supplementary Material online), were used for NU and AA sequences, respectively. Values of all parameters were calculated by RAxML. Nonparametric ML bootstrap analyses included 1,000 pseudoreplicates. For BI analyses, with incorporating the best-fit model (supplementary table S3, Supplementary Material online), two simultaneous runs

were conducted, each consisting of four chains (three heated and one cold). We allowed MrBayes to evaluate values of all parameters (e.g., setting "lset nst = 6 rates = invgamma" for the GTR + G_4 + I model). In total, chains were run for 10^6 generations, with sampling every 100 generations. The first 25% of sampled generations were discarded (burninfrac = 0.25), and the rest were used to calculate majority-rule consensus trees and posterior probabilities (PPs) for nodes. For heuristic MP analyses, we used Tree Bisection Reconnection branch swapping, with all characters equally weighted and zero-length branches collapsed. Analyses were repeated 100 times with random addition. Node support was estimated using a full heuristic bootstrap with 1,000 replicates.

To avoid using inconsistent models, the trees inferred from different codon positions and the data sets with gnetophytes removed were constructed under the same tree-building criteria and models as mentioned earlier.

Constrained Analyses

To examine whether positions of gnetophytes have effects on placements of ginkgo, ML trees was evaluated using the "multi-furcating constraint" option in RAxML. We performed a set of analyses in which each of the previously recovered positions for gnetophytes was enforced. The models used and assessment of bootstrap support (BS) values were as described previously.

Site-Stripping Test

With gnetophytes excluded, variable sites in NU data set 3 were evaluated using SlowFaster (Kostka et al. 2008). ML, BI, and MP trees were inferred from the subsets of NU data set 3 generated from progressively excluding variable sites from the highest to lowest changes (184 sites with >16 changes to 7,596 sites with >6 changes). The models used and tree-building methods were as described earlier.

Estimates of Substitution Rates and Statistical Analysis

Observed pairwise substitution rates and those estimated under the GTR + G_4 + I model were calculated using Mega 5 and RAxML, respectively. Saturation plots were then estimated using SAS v8.01 (SAS Inst., Cary, NC).

Results

Analysis of NU Sequences Results in Incongruent Placements of Ginkgo

We evaluated the effect of breadth of taxon sampling, sequence type, and tree-building method on the trees constructed with each of the three data sets (fig. 2). The sampled angiosperms, ferns, and lycophyte were used as outgroups in analyses of data sets 1, 2, and 3, respectively. For placements of ginkgo, in analyses of different NU data sets, ML and BI analyses revealed incongruent placements: The trees inferred from data set 1 supported the placement A with BS of 84% and 100% for the ML and BI trees, respectively, whereas the rest supported the placement B (BS: all > 60% for ML and PP: all > 50% for BI). However, the NU-derived MP trees congruently suggested the placement B (BS all > 85%). Apparently, in analyses of NU sequences, incongruent placements of ginkgo resulted from 1) different tree-building methods and 2) different breadth of taxon sampling (different data sets) in the ML and BI analyses. These results imply that in terms of the placement of ginkgo, the Ml and BI methods seem more sensitive to breadth of taxon sampling than the MP method.

In contrast to NU-derived trees, all AA-derived trees generated the placement B of ginkgo with strong supports (BS and PP all > 95%), regardless of which tree-building methods or data sets were used (fig. 2). However, different placements of ginkgo were found between the ML trees inferred from NU and AA sequences of data set 1. This conflict was also observed for the BI trees. Apparently, the sequence type is a determinant for the placements of ginkgo. Surprisingly, our MP trees in figure 2 congruently supported the placement B of ginkgo, suggesting that for the placement of ginkgo, the MP method is robust to the taxon sampling and sequence type.

The Position of Gnetophytes Has Influence on the Placement of Ginkgo

In all trees in figure 2, the long-branched gnetophytes were significantly supported as a monophyletic clade. Because long-branched taxa can influence placements of other ingroup taxa (Bergsten 2005), we wondered whether the placement of ginkgo was affected by gnetophytes. The "multi-furcating constraint" option of RAxML allows for initially incorporation of an incompletely constrained tree. Thus, the five previously proposed positions of gnetophytes (fig. 1) were enforced and used to assess their influences on ML analyses. However, for data set 1, the analysis of constrained position 5 did not apply because only seed plants were sampled.

Table 1 presents that the five differently constrained analyses generated only two discrepant placements of ginkgo (i.e., placements A and B). Notably, all constrained analyses of AA sequences supported the placement B with strong BS values (all > 90%). In contrast, constrained analyses of NU sequences show inconsistent placements of ginkgo, that is, the analyses of data set 1 with constrained positions 1–3 suggested the placement A (table 1), whereas the rest suggested the placement B, although two of them were weakly supported (BS < 70%). When compared with the unconstrained analysis of each data set, all analyses with constrained gnetophyte positions did not alter placements of ginkgo, except that based on NU sequences of data set 1 with the constrained position 4 (table 1). However, in terms of constrained gnetophyte positions, only the analyses with the constrained

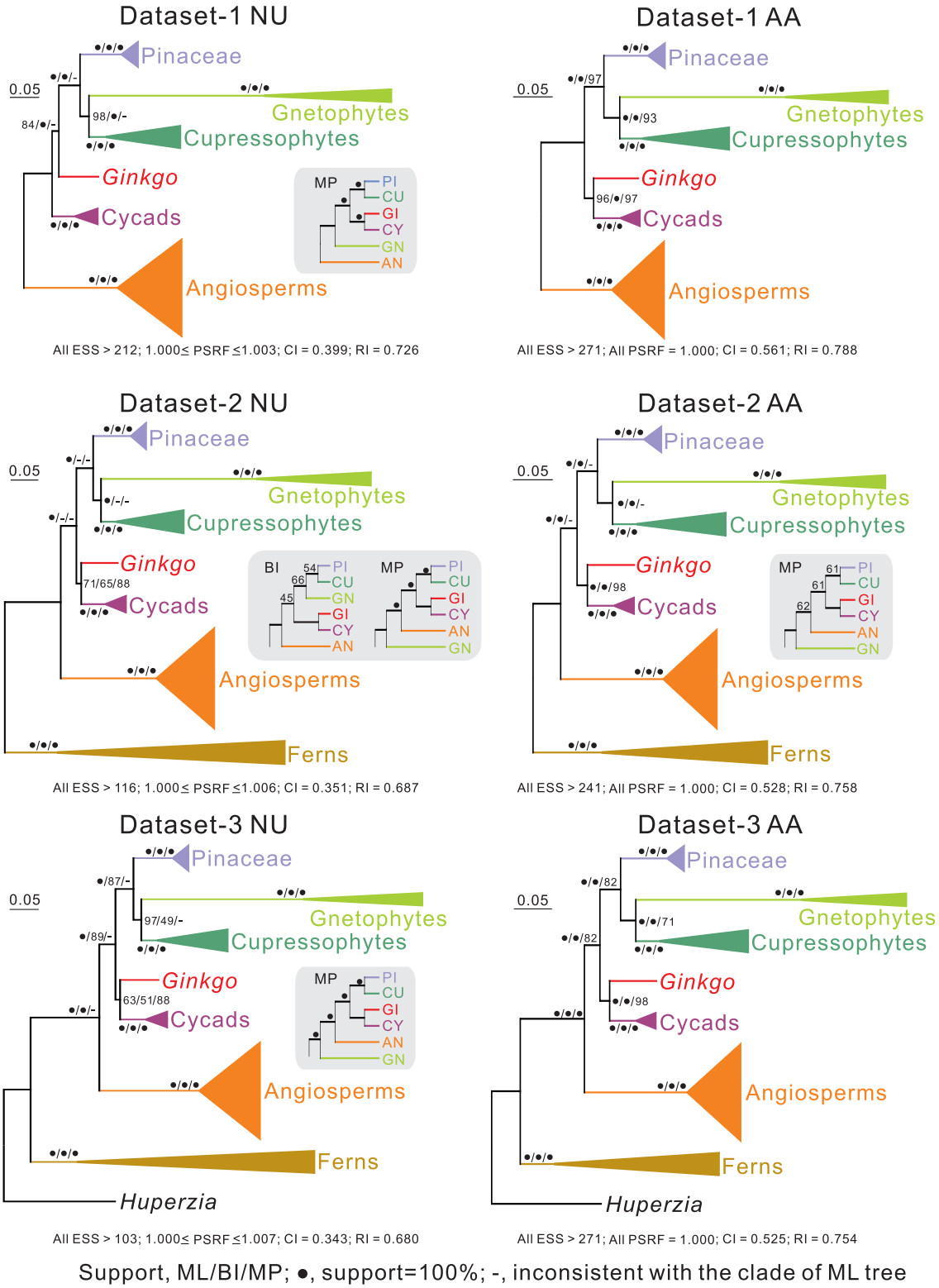

Fig. 2.—Congruent or incongruent placements of ginkgo by tree-building methods, data sets, and sequence types. Only the frameworks of ML trees are shown. Supports (%) for congruent clades are indicated for ML (1,000 bootstraps), BI (PP), and MP (1,000 bootstraps) analyses, and supports for incongruent clades are indicated in their corresponding cladograms (only topologies of six seed plant groups are shown) in gray boxes. Data sets 1, 2, and 3 include sampled taxa across seed plants, euphyllophytes (seed plants and ferns), and vascular plants, respectively. ESS, average estimated sample size; PSRF, potential scale reduction factor; CI, consistency index; RI, retention index.

Table 1

Placements of Ginkgo Inferred from Different Data Sets with Different Constrained Gnetophyte Positions

Examined Data Set	Constrained Gnetophyte Position[a]					
	Unconstrained	1	2	3	4	5
Data set 1 NU	A[b] (84)[c]	A (77)	A (75)	A (80)	B (100)	–
Data set 1 AA	B (96)	B (94)	B (98)	B (99)	B (100)	–
Data set 2 NU	B (71)	B (72)	B (82)	B (72)	B (100)	B (88)
Data set 2 AA	B (100)	B (99)	B (100)	B (96)	B (100)	B (100)
Data set 3 NU	B (63)	B (67)	B (77)	B (67)	B (100)	B (98)
Data set 3 AA	B (100)	B (99)	B (100)	B (100)	B (100)	B (100)

[a]Gnetophytes sister to 1, cupressophytes; 2, conifers; 3, Pinaceae; 4, other gymnosperms; and 5, other seed plants.
[b]Inferred placement of ginkgo: A, sister to the clade of conifers and gnetophytes and B, sister to cycads.
[c]Supported value (%) of 1,000 bootstrapping replicates.

position 4 yielded congruent placements of ginkgo with robust supports (all BS = 100%) in all data sets. Therefore, except for the position 4, constraining gnetophytes to specific positions did not affect inferred placements of ginkgo but slightly altered BS values and constraining gnetophytes as the basal-most gymnosperms (i.e., the position 4) highlighted the crucial effect of gnetophyte positions, which did alter the placement of ginkgo from A to B in analyzing NU sequences of data set 1.

The First Two and Third Codon Positions Have Conflicting Signals

To examine the influence of gnetophyte positions on phylogenetic placements of ginkgo in terms of codon positions, we further divided NU sequences of each data set into two subsets: one including the first two codon positions and the other only the third positions. From results of unconstrained ML analyses presented in table 2, signals of the first two codon positions suggested the placement B, whereas those of the third positions supported the placement A. This incongruent placement indicates conflicting signals between the first two and the third codon positions. This conflict is also revealed in the BI and MP analyses (supplementary fig. S1, Supplementary Material online). In addition, in table 2, all analyses of the first two codon positions concordantly suggested ginkgo in the placement B, regardless of which positions of gnetophytes were enforced. Thus, the placement of ginkgo inferred from the first two codon positions was unaffected by gnetophyte positions.

In contrast, in analyses of the third codon positions, results of constrained analyses appeared incongruent in the placement of ginkgo (table 2). Ginkgo was placed in the placement A with gnetophytes unconstrained or constrained at positions 1–3, whereas ginkgo was placed in the placement B in all three analyses with gnetophytes constrained at the position 4. Notably, analyses of constrained position 5 also revealed altered placements, which placed ginkgo as sister to the clade of angiosperms and cycads (placement F). Nonetheless, in the analysis of data set 3 (table 2), the BS value for the placement F was low (66%), which suggests that in terms of the influence on the placement of ginkgo, gnetophytes at the position 5 might not be as strong as at the position 4.

Removal of Gnetophytes from Data Sets Influences Placements of Ginkgo

In previous section, we demonstrated that the placement of ginkgo is associated with the position of gnetophytes in chloroplast phylogenomic estimates. As a result, we further questioned whether removal of gnetophytes from our chloroplast genome data sets would lead to congruent topologies of major seed plant lineages. Therefore, we reconstructed ML, BI, and MP trees using the same three data sets but with gnetophytes being excluded. In figure 3, all NU-derived trees placed ginkgo as sister to all conifers (placement E), whereas all AA-derived trees revealed the placement B. However, in figure 2, the NU-derived trees inferred from the data sets with gnetophytes included show incongruent placements of ginkgo. Thus, the congruent results recovered in the NU-derived trees of figure 3 suggest that inclusion of long-branched gnetophytes has influence on the placement of ginkgo. Nonetheless, this influence is not observed in the analyses of AA sequences. In all AA-derived trees, ginkgo was consistently placed in the placement B, regardless the analyzed data sets included or excluded gnetophytes (figs. 2 and 3). Of note, in figure 3, although the placement of ginkgo is consistent in the trees inferred from the same sequence type of different data sets, it is apparent that tree topologies inferred from NU and AA sequences are not conformed in the placement of ginkgo. To this end, with exclusion of gnetophytes from data sets, tree-building method and breadth of taxon sampling no longer influence the placements of ginkgo, but the sequence type is a determinant.

Tree Topology Is Misled by High Substitution Saturation of NU Sequences

To explore the possible factor leading to the incongruent placement of ginkgo in NU- and AA-derived trees in

Table 2
Placements of Ginkgo Inferred from the First Two and Third Codon Positions of the Three Data Sets with Different Constrained Gnetophyte Positions

Examined Data Set	Codon Position	Constrained Gnetophyte Position[a]					
		Unconstrained	1	2	3	4	5
Data set 1	First two	B[b] (95)[c]	B (93)	B (95)	B (93)	B (100)	–
	Third	A (99)	A (96)	A (98)	A (97)	B (81)	–
Data set 2	First two	B (100)	B (99)	B (99)	B (100)	B (100)	B (100)
	Third	A (100)	A (96)	A (99)	A (96)	B (85)	F (88)
Data set 3	First two	B (100)	B (99)	B (99)	B (99)	B (100)	B (100)
	Third	A (100)	A (96)	A (99)	A (97)	B (87)	F (66)

[a]Gnetophytes sister to 1, cupressophytes; 2, conifers; 3, Pinaceae; 4, other gymnosperms; and 5, other seed plants.
[b]Inferred placement of ginkgo: A, sister to the clade of conifers and gnetophytes; B, sister to cycads; and F, sister to the clade of angiosperms and cycads.
[c]Supported value (%) of 1,000 bootstrapping replicates.

figure 3, we conducted the site-stripping analysis. Because the topologies of three NU-derived trees shown in figure 3 are congruent, we used the data set 3 with removal of gnetophytes for tests. The monophyly of cycads was used as the reference to diagnose whether excluding variable sites from the data set also affects sampled taxa other than ginkgo. Figure 4 shows that the ML, BI, and MP trees consistently recovered the placement B of ginkgo with strong supports (all $> 85\%$) when 4,500 of the most variable sites were removed (fig. 4*A*). These variable sites are dominated by the third codon positions (79%), followed by the first (13%) and second (8%) positions (fig. 4*B*). The slopes for distribution of saturation plots changed from 0.3911 (fig. 4*C*-a) to 0.5028 (fig. 4*C*-b) (slope comparison, $P < 0.01$) after stripping the 4,500 variable sites. This sharp reduction in the level of substitution saturation strongly suggests that the placement E of ginkgo is a result from high substitution saturation in NU sequences.

Discussion

The phylogenetic placements of ginkgo were incongruent over the past two decades. Nonetheless, possible factors that lead to such incongruence have received little attention. Two well-known phylogenetic errors may be the causes: stochastic errors from analyses of insufficient data and systematic errors from misinterpretation of data (Delsuc et al. 2005). Six different placements of ginkgo (placements A–F) have been previously proposed (fig. 1). However, only two placements (placements A and B) were obtained in our analyses (fig. 2). This suggests that our chloroplast phylogenomic approach can improve the phylogenetic estimation of ginkgo, possibly because of reduced stochastic errors, a benefit of a large amount of characters used in phylogenomics (Heath et al. 2008; Lartillot and Philippe 2008).

In contrast to stochastic errors, systematic errors cannot be mitigated but may be highlighted by increasing characters in phylogenomic analyses (Delsuc et al. 2005; Jeffroy et al. 2006; Telford and Copley 2011). Signals from systematic errors can compete with historical signals, thus resulting in incongruence among independent phylogenetic analyses (Delsuc et al. 2005). Leigh et al. (2011) noted that phylogenetic congruence plays a decisive role in justifying phylogenomic analyses. Therefore, in this study, phylogenetic congruence or incongruence was used as an indicator: If an independent analysis shows phylogenetic incongruence with removal or modification of an examined factor, the factor probably contributes to the change in topology.

Broader taxon sampling was proposed as a practical avenue for accurate phylogenetic estimates (Zwickl and Hillis 2002; Dunn et al. 2008; Parfrey et al. 2010; Leigh et al. 2011). As well, increased taxon sampling should influence the inference of parameters (Pollock and Bruno 2000; Pollock et al. 2002), which is crucial for model-based methods such as the ML and BI methods. Such influence was seen in our ML and BI analyses of NU sequences. The NU-derived ML and BI trees (fig. 2) responded to increased breadth of taxon sampling and yielded altered placements of ginkgo, which implies reinterpretation of parameters when the sampled taxa were expanded from only seed plants to include also ferns. In contrast, our analysis of AA sequences generated the placement B of ginkgo in all trees with increased breadth of taxon sampling (fig. 2). These results are also in good agreement with previous studies in which all AA-derived trees supported the placement B, regardless of what breadth of taxon sampling was used (fig. 1).

Unexpectedly, the MP method is the most robust in terms of the placement of ginkgo in our study. All MP trees recovered the placement B, regardless of data set or sequence type used (fig. 2). The MP method is frequently questioned because of its high sensitivity to systematic errors such as LBA artifact (Sanderson and Shaffer 2002; Bergsten 2005; Philippe et al. 2005; Moore et al. 2011), but in certain situations, it might outperform model-based methods. All our NU-derived

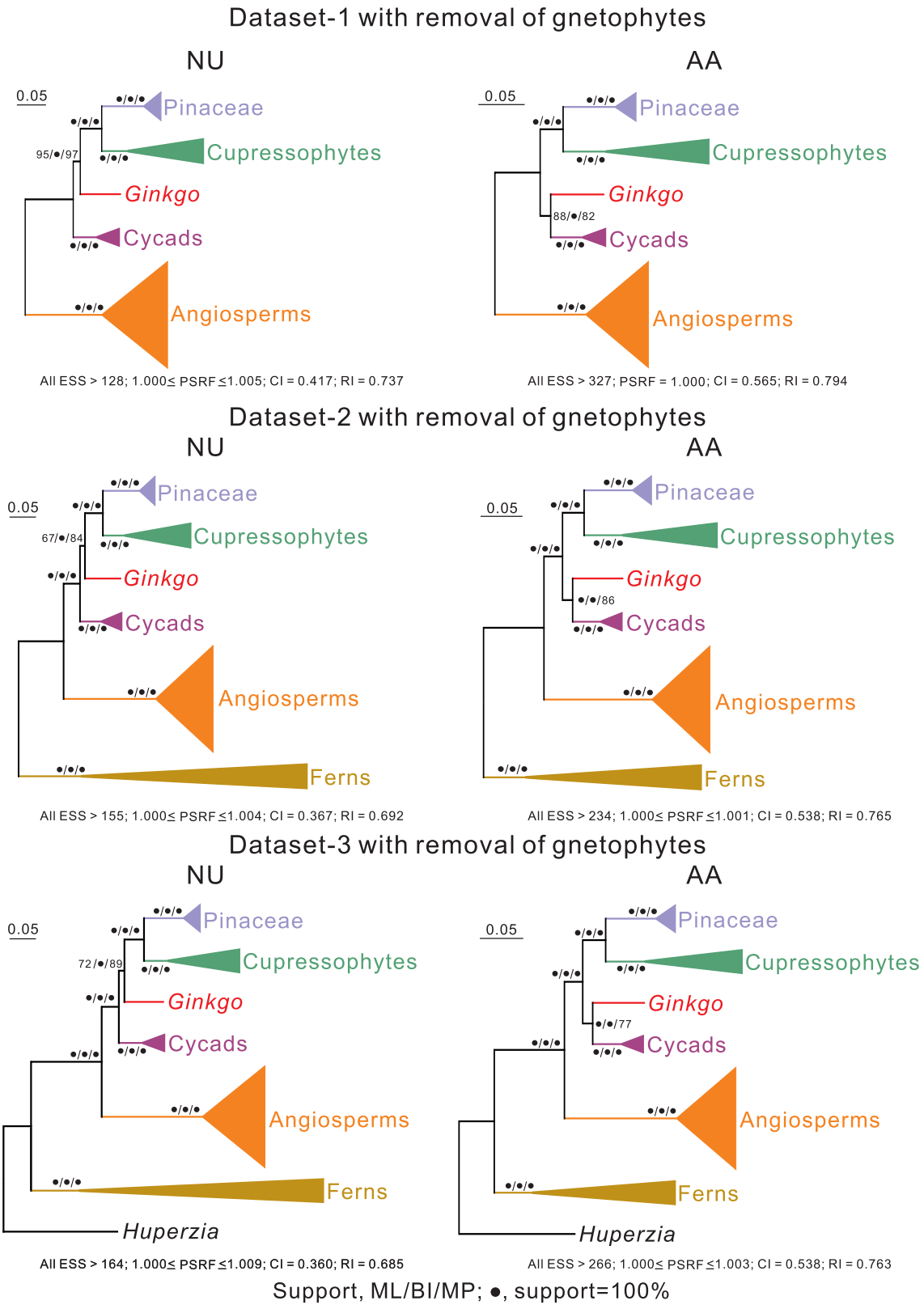

Fig. 3.—Phylogenetic trees inferred from data sets with gnetophytes excluded. The placements of ginkgo are congruent in trees inferred from the same type of sequences (NU or AA) but incongruent between NU- and AA-derived trees. Supports (%) of clades are shown in ML (1,000 bootstraps), BI (PP), and MP (1,000 bootstraps) analyses. ESS, average estimated sample size; PSRF, potential scale reduction factor; CI, consistency index; RI, retention index.

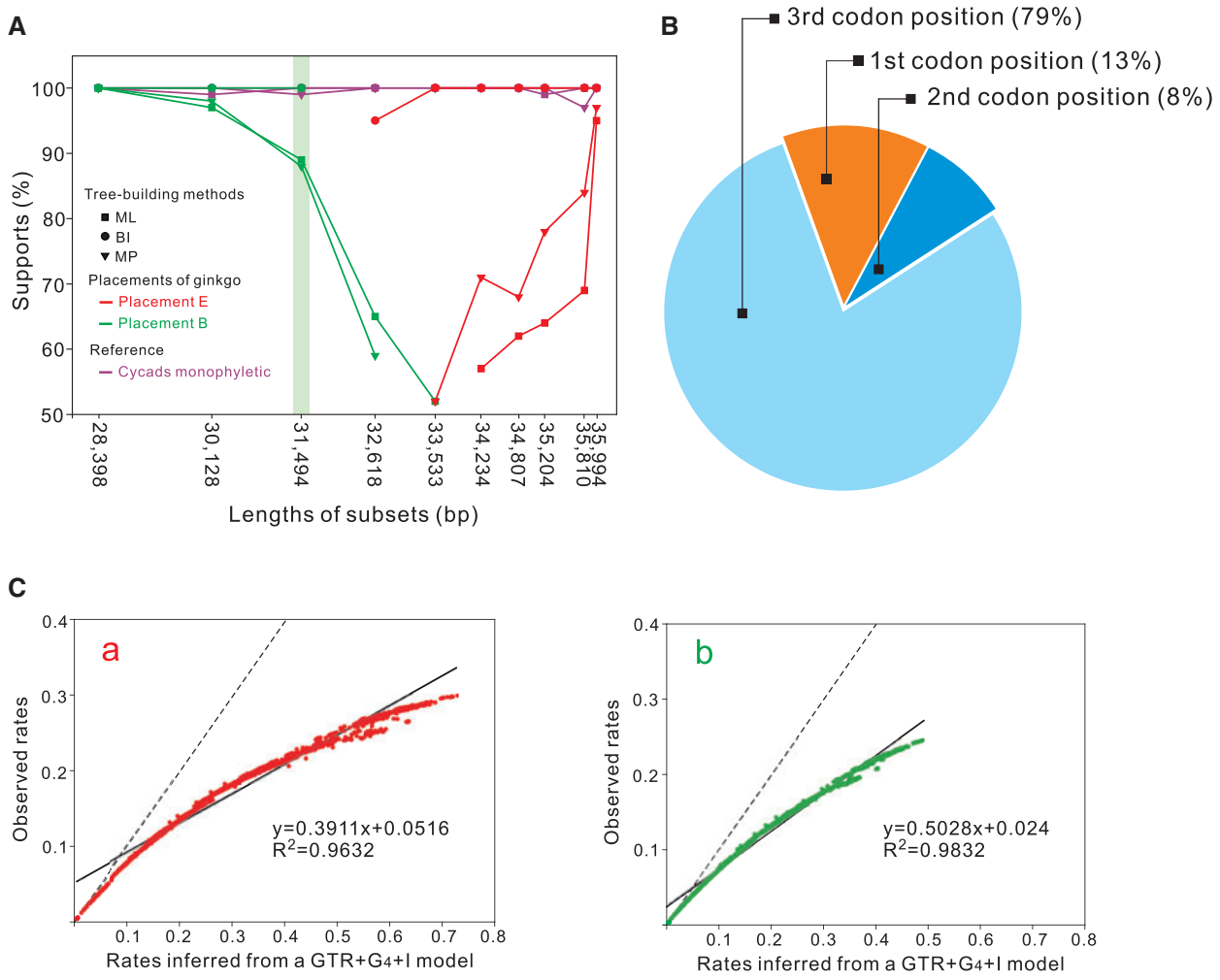

Fig. 4.—Site-stripping tests for the NU sequences of data set 3 with gnetophytes excluded. (*A*) Resulting placements of ginkgo with supports (%) in ML, BI, and MP trees inferred from the original data set (35,994 bp) and its nine subsets with gradually decreased variability. Congruent placement B of ginkgo was obtained in the three trees based on the subset (31,494 bp) that was generated by stripping 4,500 of the most variable sites from the original data set. The monophyly of cycads is shown for diagnosing the site-stripping effect upon taxa other than ginkgo. (*B*) Proportions of the first, second, and third codon positions in the 4,500 variable sites (in 100% percentage). (*C*) The levels of substitution saturation assessed before (a) and after (b) removal of the 4,500 variable sites from the original data set. Solid lines denote regression lines of saturation plots, and diagonals (dashed lines) represent no substitution saturation.

MP trees placed the long-branched gnetophytes adjacent to the outgroups (fig. 2). Such a tree topology echoes the remark of Philippe et al. (2005) that "When the two longest branches are adjacent, MP always outperforms ML."

Intriguingly, our ML analyses of NU sequences showed the effect of specific positions of gnetophytes on tree topology. Constraining gnetophytes to the position 4 consistently generated the placement B of ginkgo, even trees were inferred from the third codon positions (tables 1 and 2). Of note, except the MP tree inferred from the third codon positions of data set 1, which was likely also a result of tree topological effect mentioned earlier (position 4 of gnetophytes), all trees inferred from the third codon positions supported the placement A of ginkgo (supplementary fig. S1, Supplementary Material online). In addition, the placement of ginkgo was shifted from the placements A to F when assigning gnetophytes to the position 5 (table 2). Previously, the NU-derived ML tree of Rai et al. (2008) placed gnetophytes in the position 5 and ginkgo in the placement F (fig. 1). According to our constrained analyses, the placement F of ginkgo is likely an outcome of the specific tree topological effect.

Removal of long-branched taxa is a simple and practical strategy to detect whether taxa of interest are affected by LBA artifacts because taxa with LBA can mislead other ingroup taxa (Bergsten 2005). This strategy has been used in some plant phylogenetic studies (e.g., Chaw et al. 2000; Stefanović

et al. 2004; Karol et al. 2010). Gnetophytes have been repeatedly documented as having LBA phenomenon in phylogenetic reconstruction (e.g., Wu et al. 2007, 2011; Karol et al. 2010, Zhong et al. 2010). Indeed, after excluding gnetophytes from the data sets, all NU-derived trees are consistent in placing ginkgo in the placement E (fig. 3), which strongly indicates that the LBA of gnetophytes does influence the resulting trees inferred from NU sequences. In addition, we also demonstrated that the placement E of ginkgo (fig. 3) is due to 4,500 of the most variable sites (fig. 4*A*). These variable sites are mainly the third codon positions and contribute to substitution saturation (fig. 4*B* and *C*). Sites with high substitution saturation provide sources for misleading trees (Jeffroy et al. 2006; Rodríguez-Ezpeleta et al. 2007; Pick et al. 2010). Our results emphasize that exclusion of problematic taxa from data sets is a useful method, but sometimes this method may not lead to the correct topology.

Note that the placement A of ginkgo was retained, without any constraint, in nearly all trees inferred from the third codon positions (table 2 and supplementary fig. S1, Supplementary Material online). For deep phylogenetic analyses, the third codon positions of chloroplast genomic genes are usually saturated and must be excluded from data sets for accurate phylogenetic estimates (Chaw et al. 2004, 2005; Goremykin et al. 2004, 2009). Thus, we conclude that the placement A of ginkgo is also wrong because of high substitution saturation in the third codon positions.

In this study, we demonstrate that the factors—breadth of taxon sampling, tree-building method, position of gnetophytes, and removal of gnetophytes—do not influence the tree topologies inferred from AA sequences. Remarkably, all our AA-derived trees congruently support the placement B of ginkgo. Recent nuclear phylogenomic studies (e.g., de la Torre-Bárcena et al. 2009; Cibrián-Jaramillo et al. 2010; Finet et al. 2010; Lee et al. 2011) also consistently recovered the placement B of ginkgo (fig. 1). Although a sister relationship between ginkgo and cycads was previously reported by Goremykin et al. (1996, fig. 8) and Chaw et al. (1997) inferred from *cpITS* and 18S ribosomal DNA sequences, respectively, the former did not sample any cupressophytes and both used NJ methods. Ginkgo and cycads share a number of morphologic traits, such as haustorial pollen tubes (Friedman 1993), multiflagellated sperms (Brenner et al. 2003), simple strobili (Rudall and Bateman 2010), and some specific patterns during embryo development (Wang et al. 2011). The present chloroplast phylogenomics provides additional evidence that supports a sisterhood relationship between ginkgo and cycads.

Supplementary Material

Supplementary file S1, tables S1–S3, and figure S1 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Dr. Chung-Yen Lin and Mr. Sheng-Yao Su for assistance in computer operations, and Dr. Susanne Renner for her constructive suggestions, and greatly appreciate the two anonymous reviewers' critical reading and suggestive comments that helped improve the manuscript. This work was supported by research grants from the National Science Council, Taiwan (NSC 100-2621-B-001-003-MY3) to S.-M.C. and a regular postdoctoral fellow from Academia Sinica to C.-S.W.

Literature Cited

Bergsten J. 2005. A review of long-branch attraction. Cladistics 21: 163–193.

Bowe LM, Coat G, dePamphilis CW. 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 U S A. 97:4092–4097.

Brenner ED, Stevenson DW, Twigg RW. 2003. Cycads: evolutionary innovations and the role of plant-derived neurotoxins. Trends Plant Sci. 8: 446–452.

Burleigh JG, 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.

Chaw SM, Chang CC, Chen HL, Li WH. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol. 58:424–441.

Chaw SM, Parkinson CL, Cheng Y, Vincent TM, Palmer JD. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc Natl Acad Sci U S A. 97:4086–4091.

Chaw SM, Walters TW, Chang CC, Hu SH, Chen SH. 2005. A phylogeny of cycads (Cycadales) inferred from chloroplast *matK* gene, *trnK* intron, and nuclear rDNA ITS region. Mol Phylogenet Evol. 37: 214–234.

Chaw SM, Zharkikh A, Sung HM, Lau TC, Li WH. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol Biol Evol. 14:56–68.

Cibrián-Jaramillo A, et al. 2010. Using phylogenomic patterns and gene ontology to identify proteins of importance in plant evolution. Genome Biol Evol. 2:225–239.

Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27:1164–1165.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9:772.

de la Torre-Bárcena JE, et al. 2009. The impact of outgroup choice and missing data on major seed plant phylogenetics using genome–wide EST data. PLoS One 4:e5764.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet. 6:361–375.

Drouin G, Daoud H, Xia J. 2008. Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol. 49:827–831.

Dunn CW, et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

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

Finet C, Timme RE, Delwiche CF, Marlétaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr Biol. 20:2217–2222.

Friedman WE. 1993. The evolutionary history of the seed plant male gametophyte. Trends Ecol Evol. 8:15–21.
Goremykin V, et al. 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.
Goremykin VV, Hirsch-Ernst KI, Wölfl S, Hellwig FH. 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 21: 1445–1454.
Goremykin VV, Viola R, Hellwig FH. 2009. Removal of noisy characters from chloroplast genome–scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J Mol Evol. 68: 197–204.
Goremykin VV, et al. 2013. The evolutionary root of flowering plants. Syst Biol. 62:50–61.
Graham SW, Iles WJ. 2009. Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. Am J Bot. 96:216–227.
Hajibabaei M, Xia J, Drouin G. 2006. Seed plant phylogeny: gnetophytes are derived conifers and a sister group to Pinaceae. Mol Phylogenet Evol. 40:208–217.
Heath T, Hedtke S, Hillis D. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol. 46:239–257.
Jansen RK, et al. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci U S A. 104:19369–19374.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Karol KG, et al. 2010. Complete plastome sequences of *Equisetum arvense* and *Isoetes flaccida*: implications for phylogeny and plastid genome evolution of early land plant lineages. BMC Evol Biol. 10:321.
Kostka M, Uzlikova M, Cepicka I, Flegr J. 2008. SlowFaster, a user-friendly program for slow-fast analysis and its application on phylogeny of Blastocystis. BMC Bioinformatics 9:341.
Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Philos Trans R Soc Lond B Biol Sci. 363:1463–1472.
Lee EK, et al. 2011. A functional phylogenomic view of the seed plants. PLoS Genet. 7:e1002411.
Leigh JW, Lapointe FJ, Lopez P, Bapteste E. 2011. Evaluating phylogenetic congruence in the post-genomic era. Genome Biol Evol. 3: 571–587.
Lin CP, Huang JP, Wu CS, Hsu CY, Chaw SM. 2010. Comparative chloroplast genomics reveals the evolution of Pinaceae genera and subfamilies. Genome Biol Evol. 2:504–517.
Lin CP, Wu CS, Huang YY, Chaw SM. 2012. Complete chloroplast genome of *Ginkgo biloba* reveals the mechanism of inverted repeat contraction. Genome Biol Evol. 4:374–381.
Martin W, Deusch O, Stawski N, Grünheit N, Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10: 203–209.
Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc Natl Acad Sci U S A. 104:19363–19268.
Moore MJ, Soltis PS, Bell CD, Burleigh JG, Soltis DE. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc Natl Acad Sci U S A. 107:4623–4628.
Moore MJ, et al. 2011. Phylogenetic analysis of the plastid inverted repeat for 244 species: insights into deeper-level angiosperm relationships from a long, slowly evolving sequence region. Int J Plant Sci. 172: 541–558.
Nock CJ, et al. 2011. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J. 9:328–333.
Norstog KJ, Gifford EM, Stevenson DWM. 2004. Comparative development of the spermatozoids of cycads and *Ginkgo biloba*. Bot Rev. 70: 5–15.
Parfrey LW, et al. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. Syst Biol. 59:518–533.
Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol. 5:50.
Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Pick KS, et al. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. Mol Biol Evol. 27:1983–1987.
Pollock DD, Bruno WJ. 2000. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. Mol Biol Evol. 17:1854–1858.
Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol. 51: 664–671.
Qiu YL, et al. 2007. A non-flowering land plant phylogeny inferred from nucleotide sequences of seven chloroplast, mitochondrial and nuclear genes. Int J Plant Sci. 168:691–708.
Rai HS, Reeves PA, Peakall R, Olmstead RG, Graham SW. 2008. Inference of higher-order conifer relationships from a multi-locus plastid data set. Botany 86:658–669.
Ran JH, Gao H, Wang XQ. 2010. Fast evolution of the retroprocessed mitochondrial *rps3* gene in Conifer II and further evidence for the phylogeny of gymnosperms. Mol Phylogenet Evol. 54:136–149.
Regina TM, Quagliariello C. 2010. Lineage-specific group II intron gains and losses of the mitochondrial *rps3* gene in gymnosperms. Plant Physiol Biochem. 48:646–654.
Rodríguez-Ezpeleta N, et al. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst Biol. 56:389–399.
Ronquist F, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61: 539–542.
Rudall PJ, Bateman RM. 2010. Defining the limits of flowers: the challenge of distinguishing between the evolutionary products of simple versus compound strobili. Philos Trans R Soc Lond B Biol Sci. 365:397–409.
Rydin C, Källersjö M, Friis EM. 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 MJ, Shaffer HB. 2002. Troubleshooting molecular phylogenetic analyses. Annu Rev Ecol Syst. 33:49–72.
Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.
Stefanović S, Rice WD, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol Biol. 4:35.
Stewart CN Jr, Via LE. 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques 14: 748–751.
Swofford DL. 2002. PAUP*: phylogenetic analysis using parsimony (and other methods), 4.0 beta. Sunderland (MA): Sinauer Associates.
Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.
Telford MJ, Copley RR. 2011. Improving animal phylogenies with genomic data. Trends Genet. 27:186–195.
Wang L, et al. 2011. An embryological study and systematic significance of the primitive gymnosperm *Ginkgo biloba*. J Syst Evol. 49:353–361.
Wolfe KH, Li WH, Sharp PM. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A. 84:9054–9058.

Wu CS, Lai YT, Lin CP, Wang YN, Chaw SM. 2009. Evolution of reduced and compact chloroplast genomes (cpDNAs) in gnetophytes: selection toward a lower-cost strategy. Mol Phylogenet Evol. 52:115–124.

Wu CS, Wang YN, Hsu CY, Lin CP, Chaw SM. 2011. Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome Biol Evol. 3: 1284–1295.

Wu CS, Wang YN, Liu SM, Chaw SM. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol Biol Evol. 24:1366–1379.

Wyman SK, Jansen RK, Boore JL. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20:3252–3255.

Zhong B, Yonezawa T, Zhong Y, Hasegawa M. 2010. The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol. 27:2855–2863.

Zhong B, et al. 2011. Systematic error in seed plant phylogenomics. Genome Biol Evol. 3:1340–1348.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol. 51:588–598.

Associate editor: Geoff McFadden