The Position of Gnetales among Seed Plants: Overcoming Pitfalls of Chloroplast Phylogenomics

Bojian Zhong,*†,[1] Takahiro Yonezawa,[1] Yang Zhong,[1,2] and Masami Hasegawa*,[1,3]

[1]School of Life Sciences, Fudan University, Shanghai, China

[2]Institute of Biodiversity Science and Geobiology, Tibet University, Lhasa, China

[3]Institute of Statistical Mathematics, Tokyo, Japan

†Present address: The Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, 4442, New Zealand.

***Corresponding author:** E-mail: bjzhong@gmail.com; hasegawa@fudan.edu.cn.

Associate editor: Naruya Saitou

Research article

Abstract

The phylogenetic position of Gnetales is one of the most contentious issues in the seed plant systematics. To elucidate the Gnetales position, an improved amino acid substitution matrix was estimated based on 64 chloroplast (cp) genomes and was applied to cp genome data including all three lineages of Gnetales in maximum likelihood analyses of proteins. Although the initial analysis strongly supported the sister relation of Gnetales with *Cryptomeria* (Cupressophyta or non-Pinaceae conifers) (the "Gnecup" hypothesis), the support seems to be caused by a long-branch attraction (LBA) artifact. Indeed, by removing fastest evolving proteins that are most likely associated with the LBA, the support drastically declined. Furthermore, another analysis of partial genome data with dense taxon sampling of conifers showed that, in *psb*C, *rpl*2, and *rps*7 proteins, there are many parallel amino acid substitutions between the lineages leading to Gnetales and to *Cryptomeria*, and by further excluding these three genes, the sister relation of Gnetales with Pinaceae (the "Gnepine" hypothesis) became supported. Overall, our analyses indicate that the LBA and parallel substitutions cause a seriously biased inference of phylogenetic position of Gnetales with the cp genome data.

Key words: Gnetales, long-branch attraction, maximum likelihood, model of amino acid substitutions, parallel evolution.

Introduction

The phylogenetic position of Gnetales, a small group of gymnosperms comprised three genera (*Ephedra*, *Gnetum*, and *Welwitschia*), is one of the most controversial issues for the seed plant phylogeny (Chaw et al. 2000; Donoghue and Doyle 2000; Burleigh and Mathews 2004; Mathews 2009). From the early morphological cladistic analyses, Gnetales was considered to be sister to angiosperms (the "Anthophyte" hypothesis) (Doyle and Donoghue 1986; Rothwell and Serbet 1994). On the other hand, most molecular studies do not support the "Anthophyte" hypothesis but do not reach the agreement regarding the position of Gnetales (reviewed in Burleigh and Mathews 2004). Currently, only the three hypotheses receive some support in molecular analyses: 1) Gnetales be placed as sister to conifers as a whole (the "Gnetifer" hypothesis) (e.g., Chaw et al. 1997); 2) within conifers, and sister to Pinaceae (the "Gnepine" hypothesis) (e.g., Bowe et al. 2000; Chaw et al. 2000; Hajibabaei et al. 2006; Wu et al. 2007); and 3) within conifers, but sister to Cupressophyta (non-Pinaceae conifers) (the "Gnecup" hypothesis) (e.g., Nickrent et al. 2000; Doyle 2006; Chumley et al. 2008).

The previous analyses on the position of Gnetales with the multiple genes, however, were problematic and remained to be explored from several aspects as follows: 1) Chloroplast (cp) genomes of Gnetales and non-Pinaceae conifers were poorly sampled with respect to taxa. The genome-scale plastid data sets used previously for this issue included only one species of Gnetales, and no non-Pinaceae conifers was included (*Cycas*, *Ginkgo*, *Pinus*, and *Gnetum* in Wu et al. 2007; *Cycas*, *Ginkgo*, *Pinus*, and *Welwitschia* in McCoy et al. 2008); 2) Long-branch attraction (LBA) artifact (Felsenstein 1978; Hendy and Penny 1989; Lockhart and Steel 2005) might be operating. There are several lines of evidence of a long branch leading to Gnetales (Rai et al. 2003; Hajibabaei et al. 2006) that could potentially cause the LBA. The two widely used strategies to alleviate the LBA artifact are to remove most rapidly evolving sequences or sites (Philippe et al. 2005; Hajibabaei et al. 2006; Rai et al. 2008) and to add more taxa (Graybeal 1998; Hillis 1998; Hedtke et al. 2006; Graham and Iles 2009); and 3) Parallel substitutions can potentially cause a problem in inferring the phylogeny. Several cases of parallel substitutions, which referred to the independent evolution of similar traits, starting from the same ancestral character state, have been identified in several studies (Zhang 2006; Rokas and Carroll 2008). Such a phenomenon and convergent molecular evolution could also mislead phylogenetic inference (Christin et al. 2007; Rogozin et al. 2008; Castoe et al. 2009; Li et al. 2010; Liu et al. 2010).

With the aim of addressing these problems, we first analyzed protein-encoding genes in the currently available cp genome data set, which consists of all major groups of seed

plants, including all three genera of Gnetales (McCoy et al. 2008; Wu et al. 2009) and one representative of non-Pinaceae conifers (Hirao et al. 2008), and assessed whether the LBA artifact influences the phylogenetic placement of Gnetales. Next, we carried out several approaches to alleviate the potential biases such as the LBA and the parallel substitutions using the multiple cp genes as well as nuclear genes with more taxa from non-Pinaceae conifers.

Currently, cp genome data are frequently used in evolutionary studies of plants (e.g., Martin et al. 1998; Jansen et al. 2007; Moore et al. 2007; Zhong et al. 2009). If one is interested in resolving orders of deep branchings for the seed plants phylogeny by using cp protein-coding genes, analyses of amino acid sequences are preferable because synonymous substitutions are already saturated and accordingly should have almost no phylogenetic information, and amino acid substitutions are easier to be modeled with an empirical substitution matrix than nucleotide substitutions. For this reason, our analyses in this work are confined to the amino acid sequences of proteins. Although several empirical amino acid substitution models specific to animal mitochondrial proteins are available, only the cpREV model (Adachi et al. 2000) has been reported for plant cp proteins. With the development of the sequencing technique and the rapid accumulation of cp genome data, it seems timely to improve the cpREV model based on a much larger data set than before. A new cp amino acid substitution matrix was constructed and its performance was compared with other available matrices in this study.

Materials and Methods

Construction of a New Amino Acid Substitution Model

We retrieved 77 cp protein-encoding genes from 64 taxa, consisting of most of the major angiosperm lineages and three gymnosperms (*Cycas*, *Ginkgo*, and *Pinus*) (Jansen et al. 2007), and the tree topology in Jansen et al. (2007) was assumed. The sites that had ambiguous alignments were excluded, resulting in an alignment of 20 199 amino acid positions. Stationary frequencies of amino acids were estimated from the data, and the original cpREV model (Adachi et al. 2000) was applied for the starting parameter estimates. An empirical model of amino acid substitutions was constructed by estimating relative substitution rates between amino acids under the general time-reversible model by using PAML 4 (Yang 2007).

Data Sets for Phylogenetic Analyses

The following four data sets were used for the phylogenetic analyses.

1) Data set 1: 56 cp genes from 13 taxa

This data set comprises the 56 cp protein-encoding genes from all major groups of seed plants (11 taxa) including all three genera of Gnetales, one representative of non-Pinaceae conifers (*Cryptomeria japonica*) (only data of this species is available from non-Pinaceae conifers) and two representatives of Pinaceae conifers (*Keteleeria davidiana* and *Pinus thunbergii*), with *Physcomitrella patens* and *Marchantia polymorpha* as outgroups. The details of taxon names and GenBank accession numbers are listed in supplementary table S1, and the 56 gene names are listed in supplementary table S3 (Supplementary Material online). The data set used in estimating the new cp amino acid substitution matrix is mostly independent from Data set 1 and does not contain the phylogenetic issue going to be discussed on the position of Gnetales. Therefore, Data set 1 should be a suitable example to evaluate the performance of the new substitution model.

2) Data set 2a: 14 cp genes from 17 taxa (dense taxon sampling)

This data set includes the 14 cp protein-encoding genes collected in Graham and Iles (2009) excluding *psb*N, *ndh*B, and *ndh*F because of ambiguous alignments, with additional four taxa (*Torreya californica*, *Saxegothaea conspicua* Lindl, *Araucaria cunninghamii*, and *Cedrus deodara*) to the 13 taxa in Data set 1.

3) Data set 2b: 14 cp genes from the 13 taxa (sparse taxon sampling)

This data set is a subset of Data set 2a, with the same taxa as Data set 1.

4) Data set 3: 3 nuclear genes from 12 taxa

This data set consists of three nuclear protein-encoding genes from the 12 taxa used in Hajibabaei et al. (2006).

All the gaps and ambiguous alignments were discarded in all the data sets.

Phylogenetic Inference

Heuristic tree search with maximum likelihood (ML) phylogenetic analyses were performed by using RAxML 7.0.4 (Stamatakis 2006; Stamatakis et al. 2008) with the cpREV matrix (Adachi et al. 2000) of amino acid substitutions and a discrete gamma distribution with four rate categories. The node support was evaluated based on 100 bootstrap replications. More detailed ML analyses for a limited number of tree topologies were conducted by using CodeML of the PAML 4 package (Yang 2007). In the separate method (Kishino and Hasegawa 1989; Yang 1996), branch lengths and alpha parameters of the gamma distribution could be estimated independently for each protein, whereas in the concatenate method, the concatenated sequences were regarded as homogeneous. To examine whether the performance of phylogenetic inference changed by the improvement of the amino acid substitution model, we conducted ML analyses on the three alternative positions of Gnetales. For Data set 1, Data set 2a, and Data set 2b, the log-likelihood score of each topology was calculated with CodeML of the PAML by using the Dayhoff (Dayhoff et al. 1978), JTT (Jones et al. 1992), cpREV (Adachi et al. 2000), and the new cp substitution matrices with the discrete gamma model for site heterogeneity.

Removal of Fast-Evolving Proteins

To reduce the impact of systematic errors (Philippe et al. 2005; Nishihara et al. 2007), we detected and selectively discarded the fast-evolving proteins. For Data set 1 and 2a, the total branch lengths, base frequencies, and alpha parameters of the gamma distribution were estimated

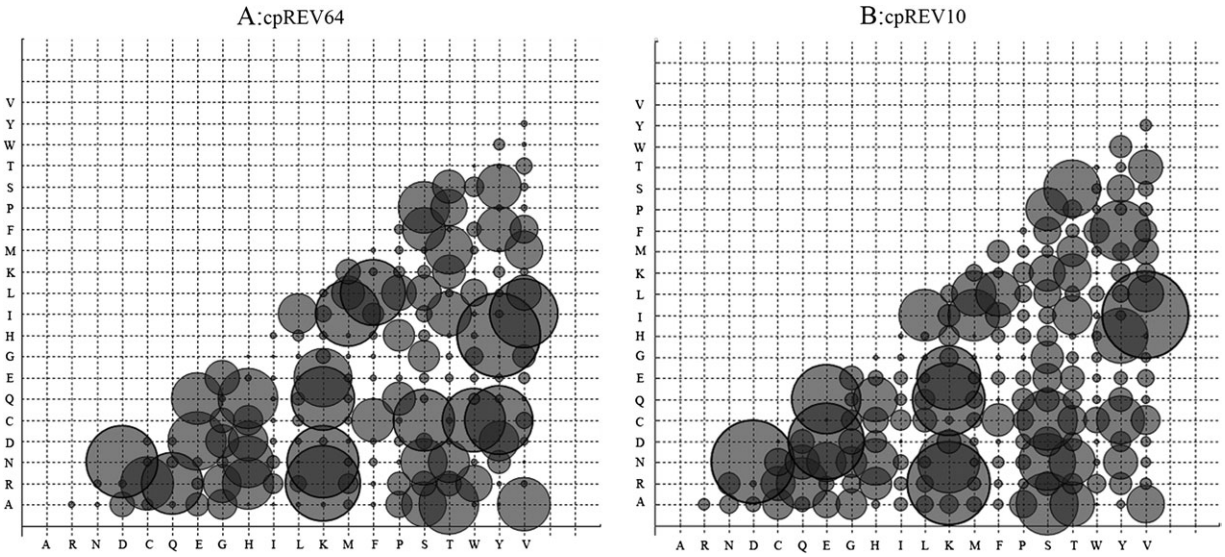

FIG. 1. The replacement rates between cpREV64 and cpREV10 matrices. Areas of bubbles are proportional to the replacement rates. Numerical values are in supplementary table S2 (Supplementary Material online).

independently for each protein by using the CodeML program in PAML 4 (Yang 2007) with our new amino acid substitution model. The proteins with long total branch lengths (larger than 2.5 substitutions per amino acid site) were excluded. The total support for a particular tree was evaluated by summing up the log-likelihood scores of each protein, and the total log-likelihood scores of the 3 candidate topologies were compared.

Reconstruction of Ancestral Character States

The ancestral site states were inferred for each interior nodes by the empirical Bayesian approach (Yang et al. 1995; Koshi and Goldstein 1996) based on densely sampled Data set 2a using the CodeML program in PAML 4 (Yang 2007) with the new amino acid substitution model. Parallel substitutions were identified by comparing ancestral and extant sequences.

Results and Discussion

The New Amino Acid Substitution Model

To improve the fitness of the cpREV model, we estimated a new cp amino acid substitution matrix using 77 protein-encoding genes from the 64 sequenced plastid genomes (Jansen et al. 2007), which is by far the most extensive data matrix applied to this issue. The new model is called the cpREV64 model, and Adachi et al. (2000) cpREV model will hereafter be referred to the cpREV10 model associated to the number of the sequences used. The replacement rates of the two cp amino acid substitution models are shown in figure 1, and the estimated amino acid substitution matrix is shown in supplementary table S2 (Supplementary Material online).

Phylogenetic Analyses and Model Comparison Based on the 56 cp Proteins from the 13 Taxa

From the ML analysis with a partitioned approach using Data set 1, the monophyly of Gnetales and *Cryptomeria* (non-Pinaceae conifers) is strongly supported (100% bootstrap support) (fig. 2). However, the branches leading to Gnetales and *Cryptomeria* are both significantly longer than those of other branches (see fig. 2 and supplementary fig. S1, Supplementary Material online), so we suspected that the Gnetales placement might reflect a LBA artifact (Felsenstein 1978; Hendy and Penny 1989).

In order to obtain a more reliable placement of Gnetales, and to evaluate the utility of various amino acid models, we applied ML analyses to compare the efficiency of amino acid models (including the cpREV64 model) with the discrete Γ distribution for site heterogeneity. Table 1 shows that the Gnetales/*Cryptomeria* clade (Tree 1: the "Gnecup" hypothesis) was strongly supported no matter which amino acid substitution model was used and that alternative trees of the Gnetales/Pinaceae clade (Tree 2: the "Gnepine" hypothesis) and the Gnetales/conifer clade (Tree 3: the "Gnetifer" hypothesis) were strongly rejected. However, there is a trend of decreasing the support of Tree 1 as the fitting of the data increases. For instance, the *P* value of the AU test (Shimodaira 2002) for Tree 2 is highest with the cpREV64+Γ model (3×10^{-4}) among those of alternative matrices, indicating that Tree 2 was relatively preferred with the improvement of the matrix. Indeed, the cpREV64+Γ model is the best model among the alternatives used in this analysis because the model gave the highest likelihood scores. It should be noted that the data set we analyzed for the phylogenetic placement of Gnetales is almost independent from the data set used in estimating the cpREV64 matrix. It may indicate that the cpREV64 matrix could relatively alleviate the LBA artifact resulting in reducing the support of Tree 1.

The phylogenomics approach using a genome-scale data is thought to be useful in estimating robust phylogenies (Philippe and Telford 2006). Several empirical studies have confirmed that the use of large amount of data could reduce the impact of stochastic error and could overcome incongruence (e.g., Bapteste et al. 2002; Philippe et al.

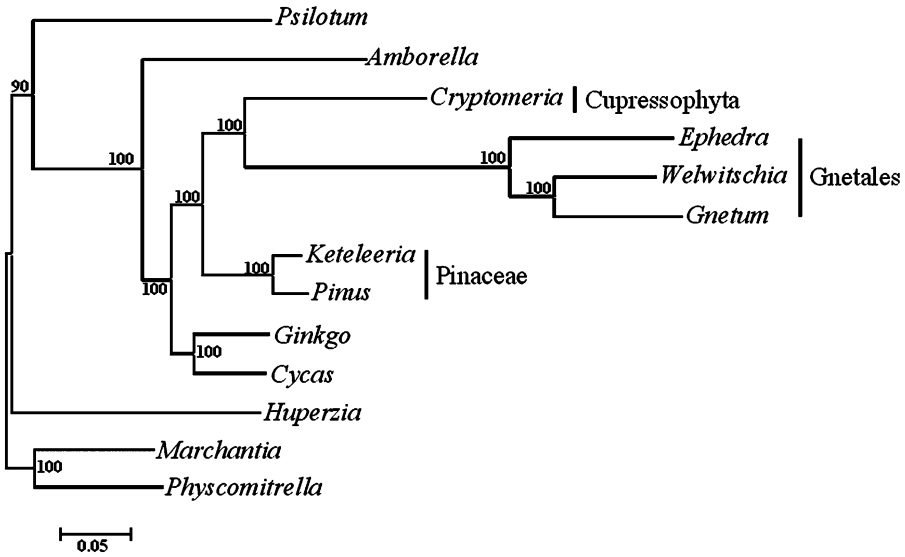

FIG. 2. The ML tree inferred from the whole cp proteins of the 13 taxa using RAxML with partitioned analysis. The numbers on the nodes indicate bootstrap probabilities. Species names: *Physcomitrella patens subsp. patens, Marchantia polymorpha, Huperzia lucidula, Psilotum nudum, Amborella trichopoda, Cycas taitungensis, Ginkgo biloba, Cryptomeria japonica, Ephedra equisetina, Gnetum parvifolium, Welwitschia mirabilis, Pinus thunbergii, Keteleeria davidiana.*

2004). However, conflicting results have also been reported, the support of erroneous trees can be enhanced by simply combining sequences (Gadagkar et al. 2005; Nishihara et al. 2007). Furthermore, the impact of systematic errors, such as the LBA artifact and the site substitution rate change over time, are amplified in the genome-scale approach. Our phylogenetic analyses of Gnetales based on the genome-scale cp data set also suggest that the LBA artifact could have biased the phylogenetic inference.

Impact of Removing Fast-Evolving Proteins for Phylogenetic Inference

It has become evident that gathering a large amount of data is not sufficient enough to obtain a reliable tree and we must try to alleviate causes that lead us to biased tree estimation. One of the well-known factors to cause biased tree estimation is the LBA artifact, and fast evolutionary rates are often associated with such a misleading effect. In order to check the possibility of the LBA artifact more in detail, we removed the 18 fastest evolving proteins (of 56 proteins) for which the total branch lengths were larger than 2.5 substitutions per site. Detailed branch lengths for each protein are shown in supplementary table S3 (Supplementary Material online). Analyses were carried out with ML and with both the concatenate and separate methods (table 2).

When all the proteins in Data set 1 were included, Tree 1 was supported with the highest log-likelihood scores. In contrast, Tree 2 could not be rejected ($P = 0.214$ by AU test) using concatenate method after removing the 18 fastest proteins (Data Set 1b). Indeed, Tree 2 became to be preferred with the highest log-likelihood scores (−48799.82) when the separate method was applied, even though Tree 1 could not be excluded. Removing the rapidly evolving sites or sequences has been suggested to improve the accuracy of phylogenetic inference (e.g., Rydin et al. 2002; Philippe et al. 2005; Rodriguez-Ezpeleta et al. 2007). Our results mentioned above may suggest that the fast-evolving proteins have caused biased estimation supporting the Gnetales/*Cryptomeria* cluster due to the LBA artifact, and the deletion of the rapidly evolving proteins has a notable effect on the phylogenetic placement of Gnetales.

Impact of Dense Taxon Sampling for Phylogenetic Placement of Gnetales

Dense taxon sampling has been proved to be another approach to break up long branches and to alleviate the impact of the LBA (e.g., Graybeal 1998; Hillis 1998; Hedtke

Table 1. Model Comparisons of the Three Alternative Placements of Gnetales Based on Data Set 1.

Models	Tree 1: Gnecup	Tree 2: Gnepine	Tree 3: Gnetifer
Dayhoff+Γ	<−127897.02>	−136.50 (<10^{-6})	−130.40 (3×10^{-5})
JTT+Γ	<−124831.23>	−119.19 (<10^{-6})	−119.62 (<10^{-6})
cpREV10+Γ	<−124646.39>	−119.42 (<10^{-6})	−115.99 (<10^{-6})
cpREV64+Γ	<−122991.77>	−103.82 (3×10^{-4})	−104.30 (<10^{-6})

NOTE.—The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown by the concatenate method. The *P* values of the AU test are given in parentheses. The following three tree topologies are examined.

Tree 1: "Gnecup" hypothesis: Gnetales sister to non-Pinaceae conifers.

Tree 2: "Gnepine" hypothesis: Gnetales sister to Pinaceae.

Tree 3: "Gnetifer" hypothesis: Gnetales sister to conifers.

Table 2. Comparison of the Log-Likelihood for the Three Alternative Trees with Concatenate and Separate Methods from the 13 Taxa.

	Concatenate Method			Separate Method		
Data Sets	**Tree 1: Gnecup**	**Tree 2: Gnepine**	**Tree 3: Gnetifer**	**Tree 1: Gnecup**	**Tree 2: Gnepine**	**Tree 3: Gnetifer**
Data set 1a: all proteins	**<−122991.77>**	**−103.82 (3×10^{-4})**	**−104.30 (<10^{-6})**	**<−118314.28>**	**−70.90 (0.003)**	**−82.11 (0.001)**
Data set 1b: fast-evolving proteins excluded	**<−50652.37>**	**−12.21 (0.214)**	**−23.37 (0.027)**	**−1.67 (0.519)**	**<−48799.82>**	**−20.54 (0.087)**
Data set 1c: *psb*C, *rpl*2 and *rps*7 further excluded	**−10.84 (0.225)**	**<−44385.20> AIC: 88856.40**	**−11.93 (0.186)**	**−25.82 (0.066)**	**<−42765.52> AIC: 88541.04**	**−19.60 (0.146)**

NOTE.—The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown. The *P* values of the AU test are given in parentheses and the Akaike information criterion (AIC) values are also shown for Data set 1c.

et al. 2006; Graham and Iles 2009). Gnetales contain only three extant genera and others are all extinct, so it is impossible to add more genera for Gnetales. In this study, we increased the number of samples from Cupressophyta (non-Pinaceae conifers) and Pinaceae groups, with three additional taxa (*T. californica*, *S. conspicua* Lindl, *A. cunninghamii*) from Cupressophyta, as well as one taxon (*C. deodara*) from Pinaceae. Although the whole cp genome data have not been sequenced from these additional taxa, a number of genes are publicly available (Graham and Iles 2009). To evaluate the performance of dense taxon sampling for the phylogenetic accuracy, the comparison of the three alternative Gnetales positions based on the dense sampling (Data sets 2a) and the sparse sampling (Data sets 2b) were conducted (table 3). When the sparsely sampled data were applied, Tree 1 was strongly supported as well as the strong rejection of Tree 2 and Tree 3. In contrast, although Tree 1 constantly had the highest log-likelihood scores, Tree 2 and Tree 3 could not be rejected (*P* value: 0.471 and 0.304, respectively, by AU test) based on the dense sampling, and the log-likelihood differences among the three alternative trees were very minor. In addition, the long branch to *Cryptomeria* in the sparse sampling was cut into short branches in the dense sampling which may have contributed to reduce the bias of tree inference (supplementary fig. S2, Supplementary Material online).

Phylogenetic Analysis Based on Nuclear Proteins

Cp genomic data have been used in most recent molecular studies to resolve the position of Gnetales (Chaw et al. 1997, 2000; Wu et al. 2007; McCoy et al. 2008), but an additional analysis of nuclear data should be helpful in clarifying this problem (Hajibabaei et al. 2006). To compare the phylogenetic position of Gnetales inferred from the nuclear genes with that from the cp genes, we conducted ML analyses based on the nuclear proteins (Data set 3). Remarkably, none of the four models supported Tree 1, and the JTT+Γ turned to be the best model (the highest log-likelihood) for the nuclear data, and preferred Tree 2 (although marginally, table 4). Therefore, combining our cp analyses with this independent evidence from the nuclear data, the "Gnecup" hypothesis (Tree 1) seems unreliable, and the "Gnepine" hypothesis (Tree 2) appears to be preferred.

Table 3. Comparison of the Log-Likelihood for the Three Alternative Trees Based on Dense Sampling and Sparse Sampling Using the cpREV64 Model.

Data Sets	**Tree 1: Gnecup**	**Tree 2: Gnepine**	**Tree 3: Gnetifer**
Data set 2b (sparse sampling)	**<−15972.28>**	**−27.90 (0.003)**	**−28.07 (0.003)**
Data set 2a (dense sampling)	**<−17193.18>**	**−1.19 (0.471)**	**−2.67 (0.304)**

NOTE.—The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown by the separate method. The *P* values of the AU test are given in parentheses.

Parallel Amino Acid Replacements and Biased Inference of the Phylogenetic Position of Gnetales

When the total branch lengths of each protein in Data set 1 were estimated, most of the proteins had a longer branch to *Cryptomeria* than to Pinaceae, but only *psb*C protein held an almost zero branch length of *Cryptomeria*, whereas the branch leading to Pinaceae had a moderate length (data not shown).

To elucidate the abnormal behavior of *psb*C protein, we estimated the ancestral site state of *psb*C and other proteins based on densely sampled Data set 2a and identified as many as 13 parallel amino acid substitutions that occurred between the branch leading to *Cryptomeria* and the common ancestral branch of Gnetales in *psb*C, *rpl*2, and *rps*7 proteins, whereas only one parallel amino acid substitution was found in *atp*B protein between the branch leading to *Pinus* and the ancestral Gnetales branch (fig. 3 and details in table 5). We suspected that the parallel substitutions may have strongly influenced the placement of Gnetales with Data set 1 (only *Cryptomeria* from Cupressophyta was included in Data set 1), the Gnetales/*Cryptomeria* grouping is likely to be preferred due to the excess of parallel amino acid changes between the branches leading to Gnetales and *Cryptomeria* in the three proteins (*psb*C, *rpl*2, and *rps*7). To reduce the bias from the parallel substitutions, we further screened out these three proteins from the data set that excluded the fast-evolving proteins and carried out ML analyses by concatenate as well as separate methods (Data Set 1c in table 2). Intriguingly, Tree 2 was favored over the alternative hypotheses with both methods, and the separate method was superior to the concatenate method based on the Akaike information criterion (Akaike 1973). Furthermore, the ML tree inference based on the Data Set 1c supported the monophyly of Gnetales and Pinaceae with a moderate bootstrap probability

Table 4. Comparison of the Log-Likelihood for the Three Alternative Trees Based on Nuclear Data (Data set 3).

	Tree 1: Gnecup	Tree 2: Gnepine	Tree 3: Gnetifer
Dayhoff+Γ	−3.04 (0.039)	<−5869.80>	−0.30 (0.529)
JTT+Γ	−3.20 (0.044)	<−5832.82>	−0.56 (0.490)
cpREV10+Γ	−3.03 (0.044)	<−5893.95>	−0.28 (0.539)
cpREV64+Γ	−3.81 (0.034)	−0.08 (0.528)	<−5961.93>

NOTE.—The log-likelihood of the ML tree (in angled brackets) and the differences in log-likelihood of alternative trees from that of the ML tree are shown by the separate method. The *P* values of the AU test are given in parentheses.

(81%) (fig. 4), even though Tree 1 could not be excluded with the 5% level of significance ($P = 0.066$ by the AU test in table 2) using the separate analysis. Additionally, the length ratio of *Cryptomeria* branch to the common ancestral branch of Pinaceae and that of the common ancestral branch of Gnetales were reduced (2.60 → 1.96 and 3.76 → 2.76, respectively, shown in supplementary figs. S1 and S3, Supplementary Material online), indicating that these three proteins that have experienced parallel evolution may have dramatically mislead the phylogenetic placement of Gnetales.

The parallel changes have been proved to bias the phylogenetic inference (Rogozin et al. 2008; Castoe et al. 2009) and the conventional phylogenetic models do not take into account the phenomenon. In this study, we reported the robust lines of evidence of parallel molecular evolution between Gnetales and *Cryptomeria* lineages in *psb*C, *rpl*2, and *rps*7 cp genes at the amino acid level, which could result in the misplacement of the phylogenetic position of Gnetales. It thus appears that after removing the fast-evolving proteins and the three parallel-evolving proteins, the remaining cp proteins tend to support Tree 2, implying that the aberrant phylogenetic signals in the cp proteins strongly support the erroneous phylogenetic placement of Gnetales.

In table S4, parallel substitutions with the ancestral Gnetales branch are shown for the branches not listed in table 5, and the numbers of parallel substitutions for each branch are shown in figure S4 under the assumption of Tree 2. The abundance of parallel substitutions is not confined to the *Cryptomeria* lineage, but is widespread in the whole Cupressophyta, whereas only one parallel substitution is found in Pinaceae. It would be reasonable to assume the four parallel substitutions in the most basal branch of non-Pinaceae conifers as shown in fig S4 could misleadingly combine Cupressophyta with Gnetales in the ML tree even with the dense sampling (Data set 2a) in table 3. It would also be reasonable to assume that some genes included in Data set 1c have also experienced parallel substitutions between ancestral Gnetales and *Cryptomeria*, and these substitutions may have prevented Data set 1c to reject Tree 1 with a high significance even using the best available model (the separate model) (table 2).

Recently, several studies demonstrated that the parallel or convergent amino acid substitutions are most likely to result from adaptive evolution (Christin et al. 2007; Castoe et al. 2009; Li et al. 2010; Liu et al. 2010). The abundant parallel substitutions between Gnetales and Cupressophyta we found in cp proteins may also be due to adaptive evolution. However, a difficulty is to explain why parallel changes are observed in various kinds of proteins including photosynthetic as well as ribosomal proteins. This situation is similar to our previous finding of supposed adaptive sites of various cp proteins in ancestral Poaceae (Zhong et al. 2009), and this problem remains to be elucidated.

Impact of Amino Acid Compositional Bias

Amino acid compositions of proteins differ among different taxa, and the difference may cause a biased estimation of the phylogeny (Lockhart et al. 1994; Foster 2004; Phillips et al. 2004). The distances of cp amino acid composition (defined by Adachi and Hasegawa 1996) are shown in table S5. For the fast + parallel genes, the distances of outgroup

Table 5. List of Parallel Amino Acid Substitutions.

Ancestral Branch to	Proteins	Parallel Substitutions with Ancestral Gnetales Branch	Site Probabilities	
Cryptomeria			C_1 → *Cryptomeria*	O → G_0
	*psb*C	10 A → T	97.2; *	100.0; 98.9
		196 V → T	99.2; *	99.7; 81.9
	*rpl*2	62 V → I	97.8; *	99.9; 95.0
	*rps*7	7 A → E	99.0; *	100.0; 99.7
		87 R → K	91.8; *	99.8; 99.2
		101 G → E	99.1; *	100; 99.8
		138 R → K	95.0; *	99.9; 99.6
Torreya/*Cryptomeria*			C_0 → C_1	O → G_0
	*Psb*C	173 V→ I	99.7; 98.8	99.7; 98.5
	*rpl*2	85 M → I	99.6; 99.0	100; 96.4
		156 G → S	97.8; 99.8	98.8; 71.1
	*rps*7	9 S → F	99.4; 99.7	99.9; 94.7
		68 V → L	99.7; 99.5	100; 99.9
		69 T → I	99.4; 99.0	100; 99.8
Pinus			P_1 → *Pinus*	O → G_0
	*atp*B	65 S → P	99.6; *	99.5; 84.0
Keteleeria	None	None	None	None
Pinus/*Keteleeria*	None	None	None	None

NOTE.—Only branches leading to the species used in the cp genome analysis are shown. A list for other branches is given in supplemental table S4 (Supplementary Material online). Branch labels refer to those in figure 3. The probabilities of ancestral nodes are shown in % (* refers to a terminal). The numbering of amino acid sites experiencing parallel substitutions are those of *Pinus thunbergii* (Wakasugi et al.1994).

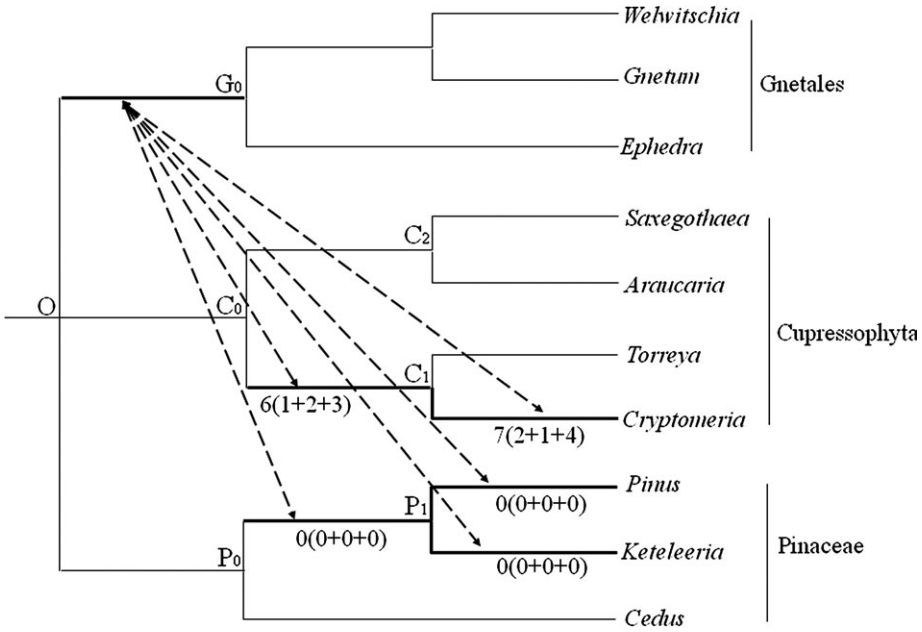

Fig. 3. Parallel amino acid replacements between the branches leading to *Cryptomeria* and *Pinus*–*Keteleeria* with the ancestral Gnetales branch. Numbers of parallel sites were indicated, for example, 7 (2 + 1 + 4) means that 2, 1, and 4 sites involved in parallel of *psb*C, *rpl*2, and *rps*7 proteins, respectively, (7 in total) in the terminal branch of *Cryptomeria*. More details are shown in table 5.

species (*Amborella*, *Cycas*, and *Ginkgo*) to Gnetales are much longer than those to Pinaceae (*Pinus* and *Keteleeria*), and the distance to Cupressophyta (*Cryptomeria*) is intermediate. This trend is held for Data set 1c (exclusion of fast + parallel genes), although in a more moderate way. The extreme difference of amino acid compositions in Gnetales from those in Pinaceae is likely to be one of the reasons why Data set 1a gave a strong support for Tree 1. Furthermore, the retention of the same trend even in a more moderate way after excluding the fast + parallel genes may partially explain why Data set 1c cannot exclude Tree 1 with a high significance.

Impact of Alternative Outgroup Relationship of *Cycas* and *Ginkgo*

Cycas and *Ginkgo* are strongly monophyletic from the ML tree (fig. 2), and therefore we assumed this clade in the subsequent analyses. However, several molecular analyses suggest that *Cycas* is the basal Gymnosperm lineage (e.g., Bowe et al. 2000; Chaw et al. 2000). To check whether the alternative relationship between *Cycas* and *Ginkgo* could affect the position of Gnetales, we further compared the log-likelihood of the three alternative positions of Gnetales with *Cycas* as the basal Gymnosperm for the three data sets (table S6). The results indicate that "Gnepine" tree is favored by excluding fast-evolving and parallel genes, which are congruent with the analyses assumed the monophyly of *Cycas* and *Ginkgo*, and the relative supports of the three alternative trees remain essentially unchanged from those of table 2. So the support of "Gnepine" tree is robust even though alternative relationship between *Cycas* and *Ginkgo* is assumed.

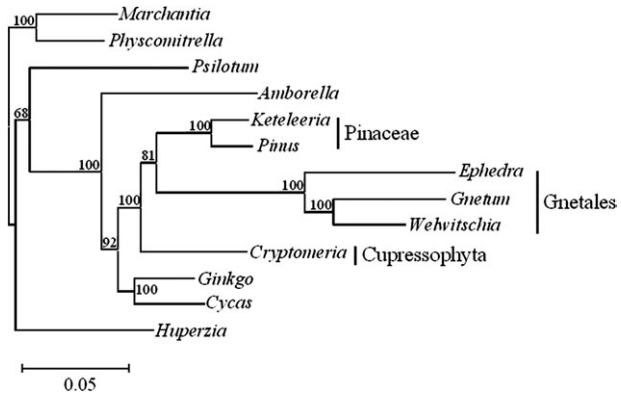

Fig. 4. The ML tree inferred from Data set 1c using RAxML with partitioned analysis. The numbers on the nodes indicate bootstrap probabilities.

Conclusion

Phylogenetic position of Gnetales has been a mysterious issue in the seed plants phylogenetics. With the cp amino acid substitution matrix based on the largest data set for this purpose, we are able to improve the current substitution model to help resolve this controversial problem in this study. We found that the LBA artifact and the parallel changes play significant roles in misleading the phylogenetic placement of Gnetales. In particular, our analyses suggest that parallel amino acid changes have a strong impact in grouping Gnetales with *Cryptomeria* (the "Gnecup" hypothesis). Here, we have demonstrated that removing fast-evolving genes and increasing taxon sampling can effectively alleviate the LBA artifact, thereby recovering a sister-group position of Gnetales to Pinaceae (the "Gnepine" hypothesis). Furthermore, this hypothesis is increasingly supported by removing the parallel-evolving proteins. Additionally, independent lines of evidence supporting the "Gnepine" hypothesis were recently provided by plastid

structural losses, such as the loss of all *ndh* genes (Braukmann et al. 2009) and the loss of *rps*16 (Wu et al. 2007, 2009) commonly observed between Gnetales and Pinaceae lineages.

Therefore, although the aberrant phylogenetic signals in our data sets prevented us to draw a firm conclusion on the position of Gnetales, the congruence of our analyses makes us confident in supporting the "Gnepine" hypothesis. Currently, cp genome data are available from only one species (*C. japonica*) among non-Pinaceae conifers. Additional cp genomic data of the non-Pinaceae conifers is expected to shed more light on the position of Gnetales.

Supplementary Material

Supplementary tables S1–S6 and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank David Penny for helpful comments. We also thank the associate editor and two anonymous reviewers for constructive suggestions on the manuscript. This work was supported by National Science Foundation of China (30925004).

References

Adachi J, Hasegawa M. 1996. MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Compt Sci Monogr Inst Statist Math Tokyo*. 28:1–150.

Adachi J, Waddell PJ, Martin W, Hasegawa M. 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J Mol Evol*. 50:348–358.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second International Symposium on Information Theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.

Bapteste E, Brinkmann H, Lee JA, et al. (11 co-authors). 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proc Natl Acad Sci U S A*. 99:1414–1419.

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.

Braukmann TW, Kuzmina M, Stefanovic S. 2009. Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. *Curr Genet*. 55:323–337.

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.

Castoe TA, de Koning AP, Kim HM, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A*. 106:8986–8991.

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, 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.

Christin PA, Salamin N, Savolainen V, Duvall MR, Besnard G. 2007. C_4 photosynthesis evolved in grasses via parallel adaptive genetic changes. *Curr Biol*. 17:1241–1247.

Chumley TW, McCoy SKR, Raubeson LA. 2008. Gnedeep: Exploring Gnetalean affinities in seed plant phylogeny with 83 plastid genes. Botany 2008: Joint Annual Meeting of Canadian Botanical AssociationAmerican Fern Society, American Society of Plant Taxonomists, and the Botanical Society of America. Vancouver, British Columbia (Canada). Available form: http://2008.botany-conference.org/engine/search/index.php?func=detail&aid=770.

Dayhoff MO, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff MO, editor. Atlas of protein sequence and structure. Washington (DC): National Biomedical Research Foundation. Vol. 5. p. 345–352.

Donoghue MJ, Doyle JA. 2000. Seed plant phylogeny: demise of the anthophyte hypothesis. *Curr Biol*. 10:R106–R109.

Doyle JA. 2006. Seed ferns and the origin of angiosperms. *J Torrey Bot Soc*. 133:169–209.

Doyle JA, Donoghue MJ. 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 and compatibility methods will be positively misleading. *Syst Zool*. 27:401–410.

Foster PG. 2004. Modeling compositional heterogeneity. *Syst Biol*. 53:485–495.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *J Exp Zool B Mol Dev Evol*. 304:64–74.

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

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol*. 47:9–17.

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.

Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol*. 55:522–529.

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

Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst Biol*. 47:3–8.

Hirao T, Watanabe A, Kurita M, Kondo T. 2008. Complete nucleotide sequence of the *Cryptomeria japonica* D. Don. chloroplast genome and comparative chloroplast genomics: diversified genomic structure of coniferous species. *BMC Plant Biol*. 8:70.

Jansen RK, Cai Z, Raubeson LA, et al. (16 co-authors). 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.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *CABIOS*. 8:275–282.

Kishino H, Hasegawa M. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*. 29:170–179.

Koshi JM, Goldstein RA. 1996. Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol*. 42:313–320.

Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene *Prestin* unites echolocating bats and whales. *Curr Biol*. 20:R55–R56.

Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol*. 20:R53–R54.

Lockhart PJ, Steel MA, Hendy MD, Penny D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol*. 11:605–612.

Lockhart PJ, Steel MA. 2005. A tale of two processes. *Syst Biol*. 54:948–951.

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393:162–165.

Mathews S. 2009. Phylogenetic relationships among seed plants: persistent questions and the limits of molecular data. *Am J Bot*. 96:228–236.

McCoy SR, Kuehl JV, Boore JL, Raubeson LA. 2008. The complete plastid genome sequence of *Welwitschia mirabilis*: an unusually compact plastome with accelerated divergence rates. *BMC Evol Biol*. 8:130.

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*. 104:19363–19368.

Nickrent DL, Parkinson CL, Palmer JD, DuV RJ. 2000. Multigene phylogeny of land plants with special reference to bryophytes and the earliest land plants. *Mol Biol Evol*. 17:1885–1895.

Nishihara H, Okada N, Hasegawa M. 2007. Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol*. 8:R199.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst*. 36:541–562.

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PW, Casane D. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol*. 21:1740–1752.

Philippe H, Telford M. 2006. Large-scale sequencing and the new animal phylogeny. *Trends Ecol Evol*. 21:614–620.

Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. *Mol Biol Evol*. 21: 1455–1458.

Rai HS, O'Brien HE, Reeves PA, Olmstead RG, Graham SW. 2003. Inference of higher-order relationships in the cycads from a large chloroplast data set. *Mol Phylogenet Evol*. 29:350–359.

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.

Rodriguez-Ezpeleta N, Brinkmann H, Roure B, Lartillot N, Lang BF, Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst Biol*. 56:389–399.

Rogozin IB, Thomson K, Csuros M, Carmel L, Koonin EV. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. *Biol Direct*. 3:7.

Rokas A, Carroll SB. 2008. Frequent and widespread parallel evolution of protein sequences. *Mol Biol Evol*. 25:1943–1953.

Rothwell GW, Serbet R. 1994. Lignophyte phylogeny and the evolution of spermatophytes: a numerical cladistic analysis. *Syst Bot*. 19:443–482.

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.

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

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol*. 57:758–771.

Wakasugi T, Tsudzuki J, Ito S, Nakashima K, Tsudzuki T, Sugiura M. 1994. Loss of all *ndh* genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc Natl Acad Sci U S A*. 91:9794–9798.

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

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.

Yang Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J Mol Evol*. 42:587–596.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141:1641–1650.

Zhang J. 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nat Genet*. 38:819–823.

Zhong BJ, Yonezawa T, Zhong Y, Hasegawa M. 2009. Episodic evolution and adaptation of chloroplast genomes in ancestral grasses. *PLoS ONE*. 4:e5297.