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

Chung-Shien Wu,† Ya-Nan Wang,† Shu-Mei Liu,* and Shu-Miaw Chaw**

*Research Center for Biodiversity, Academia Sinica, Taipei, Taiwan; and †School of Forestry and Resource Conservation, National Taiwan University, Taipei, Taiwan

Phylogenetic relationships among the 5 groups of extant seed plants are presently unsettled. To reexamine this long-standing debate, we determine the complete chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 protein-coding genes encoded in the cpDNA of *Gnetum parvifolium*. The cpDNA of *Cycas* is a circular molecule of 163,403 bp with 2 typical large inverted repeats (IRs) of 25,074 bp each. We inferred phylogenetic relationships among major seed plant lineages using concatenated 56 protein-coding genes in 37 land plants. Phylogenies, generated by the use of 3 independent methods, provide concordant and robust support for the monophylies of extant seed plants, gymnosperms, and angiosperms. Within the modern gymnosperms are 2 highly supported sister clades: *Cycas–Ginkgo* and *Gnetum–Pinus*. This result agrees with both the "gnetifer" and "gnepines" hypotheses. The sister relationships in *Cycas–Ginkgo* and *Gnetum–Pinus* clades are further reinforced by cpDNA structural evidence. Branch lengths of *Cycas–Ginkgo* and *Gnetum* were consistently the shortest and the longest, respectively, in all separate analyses. However, the *Gnetum* relative rate test revealed this tendency only for the 3rd codon positions and the transversional sites of the first 2 codon positions. A Ψ*tufA* located between *psbE* and *petL* genes is here first detected in *Anthoceros* (a hornwort), cycads, and *Ginkgo*. We demonstrate that the Ψ*tufA* is a footprint descended from the chloroplast *tufA* of green algae. The duplication of *ycf2* genes and their shift into IRs should have taken place at least in the common ancestor of seed plants more than 300 MYA, and the *tRNAPro-GGG* gene was lost from the angiosperm lineage at least 150 MYA. Additionally, from cpDNA structural comparison, we propose an alternative model for the loss of large IR regions in black pine. More cpDNA data from non-Pinaceae conifers are necessary to justify whether the gnetifer or gnepines hypothesis is valid and to generate solid structural evidence for the monophyly of extant gymnosperms.

Introduction

The Cycadales (cycads), comprising about 300 species, have survived since the Pennsylvanian era, approximately 300 MYA (Norstog and Nicholls 1997; Hill et al. 2003). They dominated the Mesozoic forests along with conifers and ginkgos. Cycads are considered to be closely linked with spore-producing ferns because their trunks bear pinnately compound leaves and lack axillary buds, but have unique girdling leaf traces and dichotomous branching (vs. axillary branching in other seed plants), which occur in some ferns but in no other seed plants (Stevenson 1990). Additionally, among 4 extant gymnosperm groups (cycads, conifers, *Ginkgo*, and Gnetales), the cycads are considered relatively ancient because their mature pollen has multiciliate sperms and ovules are borne on the margins of leaf-like megasporophylls (Stevenson 1990). Because of their old fossil records and primitive morphology, cycads are traditionally treated as the oldest or the basal-most lineage among the 5 living groups of seed plants (Gymnosperms: Cycadales, Coniferales, Ginkgoales, and Gnetales; and angiosperms) (Brenner et al. 2003).

In the past decade, molecular data have been widely used to reexamine the traditional evolutionary schemes of seed plants but have generated an even more diverse set of phylogenetic hypotheses, especially about relationships among the 4 surviving groups of gymnosperms and the angiosperms (see also review by Burleigh and Mathews 2004). For example, some authors (Hamby and Zimmer 1992; Albert et al. 1994; Rydin et al. 2002; Schmidt and Schneider-Poetsch 2002) have placed Gnetales (including Gnetaceae, Ephedraceae, and Welwitschiaceae) as the deepest diverging lineage in seed plant evolution, which implies that the divergence of Gnetales predates other gymnosperms and that the extant 4 gymnosperm lineages are paraphyletic. In contrast, molecular phylogenetic trees reconstructed by others (Goremykin et al. [1996]; Chaw et al. [1997]; Bowe et al. [2000]; Chaw et al. [2000]) congruently suggested that the extant 4 gymnosperm orders are monophyletic, and none is a sister group of the angiosperms. However, these studies indicate different relationships between the 4 extant gymnosperm lineages. Bowe et al. (2000) and Chaw et al. (2000) even resolved Gnetales as a sister clade to the Pinaceae, which is named the "gnepine" hypothesis by the latter authors and suggests that the conifer families are polyphyletic. Recently, Wang (2004) reported a new Permian Gnetalean cone (ca. 270 MYA), which provides unequivocal evidence that the Gnetales coexisted closely with the conifers since the Paleozoic and satisfies "a precondition for establishing monophyletic gymnosperms (Wang 2004)."

However, seed plant phylogenies inferred from molecular data have also been complicated by the long-branch attraction (LBA) phenomenon (Sanderson et al. 2000; Magallón and Sanderson 2002; Rydin et al. 2002; Rai et al. 2003; Hajibabaei et al. 2006) inherited in the Gnetales. The reason to the faster substitution rates of Gnetum remained to be explored. However, taxa with faster evolving sequences are claimed to tend to cluster together in a phylogenetic tree (Felsenstein 1978; Huelsenbeck 1995; Bergsten 2005). As a result, the tree topology might be biased; hence, previously published phylogenetic trees that included the Gnetales were often questioned. The most widely suggested remedy for LBA artifacts is adding more

Key words: *Cycas taitungensis*, *Gnetum parvifolium*, *Ginkgo*, chloroplast genome, gymnosperms, seed plants, phylogeny, evolution, substitution rate.

E-mail: smchaw@sinica.edu.tw.

Mol. Biol. Evol. 24(6):1366–1379. 2007
doi:10.1093/molbev/msm059
Advance Access publication March 22, 2007

taxa (Swofford et al. 1996; Graybeal 1998; Hillis 1998) or using slowly evolving genes or sites in a phylogenetic analysis (Aguinaldo et al. 1997; Hajibabaei et al. 2006).

The cpDNA sequences are useful for resolving the plant phylogeny at deep levels of evolution because of their lower rates of silent nucleotide substitution (see review by Raubeson and Jansen 2005). Additionally, concatenating sequences from many genes may overcome the problem of multiple substitutions that cause the loss of phylogenetic information between cp lineages (Lockhart et al. 1999). Furthermore, structural characters in cpDNAs, such as gene order/segment inversions, expansion/contraction of the inverted repeat (IR) regions, and loss/gain of genes, can serve as powerful markers for phylogenetic inference (Raubeson and Jansen 2005). For example, an inversion flanking the *petN* and *ycf2* genes was found to occur in all cpDNAs of vascular plants except lycopods (Lycopodiopsida), which suggests that lycopsids are the basal-most lineage of vascular plants (Raubeson and Jansen 1992b); an unusual loss of *rpl21* in the cpDNAs of seed plants supports the monophyly of the group (Gallois et al. 2001); and a common duplication of the *trnH–rps19* gene cluster in IRs distinguishes the monocots from the dicots (Chang et al. 2006).

To resolve the long-standing controversies over the phylogenies of the 5 major extant seed plant lineages and to gain more insights into the diversity and evolution of cpDNA structures in seed plants, we determined the whole cpDNA sequence of the first cycad, *Cycas taitungensis*, and 56 cp protein–coding genes of a Gnetales representative, *Gnetum parvifolium*. The 56 genes are common to the other known land plants. A data set concatenating the corresponding 56 protein-coding genes from the present 2 gymnosperms and other 35 cpDNAs of land plants, including 2 bryophytes, 1 lycophyte, 1 fern, 3 gymnosperms, and 28 angiosperms (supplementary table 3, Supplementary Material online), was analyzed by 3 phylogenetic methods. The sequence data set is by far the largest with the most diverse seed plant taxa sampled. Relationships of *Cycas*, *Gnetum*, and other seed plant lineages were inferred and discussed. Additionally, structural organizations of the 37 cpDNAs sampled were critically compared and informative changes were utilized to test phylogenetic inferences resulting from the present data set as well as prior studies.

Materials and Methods

cpDNA Extraction of *C. taitungensis* and Sequencing

We used an 8-year-old *C. taitungensis* tree grown in the greenhouse of Academia Sinica as material. Young leaves less than 10 days old were collected for isolating intact chloroplasts, which were fractionated with step percoll (30−50%) gradient (Robinson and Downton 1984). DNAaseI was used to digest contaminated nuclear genome (nrDNA) included in the fractionated intact chloroplasts. The cpDNA was isolated according to a CTAB-based protocol (Stewart and Via 1993) and sheared into random fragments of 2–3 kbp by use of a Hydroshear device (Genomic Solutions Inc., Ann Arbor, MN) and then cloned into the pBluescriptSK vector to generate a shotgun library. Shotgun clones were propagated, and their plasmids were used as templates for subsequent sequencing. The sequencing reaction involved use of the BigDye terminator cycle sequencing kit (Applied Biosystems, Foster City, CA) according to the manufacturer's protocol. The DNA sequencer was an Applied Biosystems ABI 3700. The sequences determined from both ends of each shotgun clone were accumulated, trimmed, aligned, and assembled by use of the Phred-Phrap programs (Phil Green, University of Washington, Seattle, WA). Each nucleotide has about 8 × coverage. Gaps were filled with specific primers.

Amplification and Sequencing of *G. parvifolium* cpDNA

Young leaves of *G. parvifolium* were harvested from a plant growing in the greenhouse of Academia Sinica. Total DNA was extracted as above. The cpDNA fragments were amplified by use of long-range polymerase chain reaction (PCR) (TaKaRa LA Taq, Takara Bio Inc, Shiga, Japan) with primers designed according to the conserved regions of known cpDNA sequences. We covered the entire *G. parvifolium* cpDNA with 10 partially overlapped PCR fragments, which are approximately 7–16 kb in length. Each fragment was sequenced by combining at least 2 independent PCR amplicons. The IR regions of the cpDNA were amplified separately, each with 2 PCR amplicons extending to the large and single-copy regions, and overlapped in the middle of the respective repeats. PCR products were purified and eluted by electrophoresis with low-melting agarose. Eluted DNA fragments were hydrosheared, cloned, sequenced, and assembled with the same methods as for *C. taitungensis*.

Gene Annotation and Repeat Sequence Analysis

The obtained cpDNA sequences of the 2 species were annotated by use of DOGMA (Dual Organellar GenoMe Annotator) (Wyman et al. 2004). For genes with low sequence identity, a manual annotation was performed. We first identified the position of the start and stop codons, then translated the prospective genes into amino acid sequences by use of the standard/bacterial code. Analysis of repeat sequence was carried out by use of REPuter (http://www.genomes.de/). The settings for identification of direct repeat and IRs included a size more than 15 bp and a Hamming distance of 3. Low complexity and nested repeats were ignored.

Sequencing of Chloroplast *Ycf2* and *TufA* Genes in Other Cycads and *Ginkgo*

Total DNA of other cycads and *Ginkgo biloba* from leaves were extracted by use of the above method. The primer pair, petG-F (5′-ATAGGAATTAGACCTAACACAATTCC-3′) and psbE-R (5′-CTATGGATAACCCAGTATCGAATACT-3′), for amplifying the spacer region between *petG* and *psbE* genes and rpl32-F (5′-ATGGCAGTTCCGAAGAAACG-3′) and ccsA-R (5′-TGCGCCTAACCACCGATTAT-3′) for amplifying the region between *rpl32* and *ccsA* genes, resulted in 2- to 2.5-kb fragments under the following reaction conditions: 94 °C for 5 min, followed by 30 cycles of 94 °C for 20 s, 58 °C

for 30 s, and 72 °C for 2 min, ending with a 7-min extension at 72 °C. For verification of the presence of *ycf2* genes in the IRs of the cpDNAs of *G. parvifolium* and *G. biloba*, the reverse primer ycf2-R (5′-TTTKACRGGATTCARCCARTTGTC-3′) was designed from the conserved region of *ycf2* to combine with 1 of the 2 forward primers, rps19-F (5′-CAATTTGTGDYCTACCATACGATC-3′) and trnK-F (5′-GGGTTGCTAACTCAAYGGTAGAG-3′), on the basis of the sequences of *rps19* and *trnK* genes, respectively. The *rps19-ycf2* and *trnK-ycf2* fragments were amplified by long-range PCR. The PCR products were purified by use of the Gel Extraction System (Gel-M, Viogene Inc., Taiwan) and directly sequenced as described above.

Reverse Transcriptase–Polymerase Chain Reaction

To verify the expression of orf75, total RNAs were extracted from the young leaves of *C. taitungensis* with use of Trizol (Invitrogen, Carlsbad, CA). For reverse transcriptase–polymerase chain reaction (RT-PCR) assay, total RNA was treated with DNAaseI and then extracted with phenol–chloroform to eliminate any DNA contamination. The resulting RNA was reverse transcribed to make cDNA with a gene-specific primer orf75-R (5′-TCCGATCTCTACGCATTTCA-3′) and the Superscript II RT (Invitrogen, Indianapolis, IN). The primer pair orf75-F (5′-TCCGATCTCTACGCATTT-3′) and orf75-R amplified a 208-bp fragment under the following reaction conditions: 94 °C for 2 min; followed by 30 cycles of 94 °C for 20 s, 55 °C for 30 s, and 72 °C for 30 s; and ending with a 4-min extension at 72 °C.

Sequence Alignment and Phylogenetic Analysis

We extracted 56 plastid protein–coding genes (supplementary table 4, Supplementary Material online) from 35 land plants in the National Center for Biotechnology Information (NCBI) data bank (supplementary table 3, Supplementary Material online) and the 2 taxa sequenced in the present study. Alignment of each gene first involved GENEDOC 2.6 (Nicholas KB and Nicholas HB 1997) followed by a manual adjustment and then concatenation as a sequence data set for calculating distances, substitution rate comparison, and phylogenetic analysis. The Kimura 2-parameter (K2P) model (Kimura 1980) implemented in the software program MEGA3 (Kumar et al. 2004) was used to estimate pairwise transitional (Ts) and tranversional (Tv) distances among taxa. In phylogenetic tree reconstruction, we applied 3 methods: Bayesian inference (BI), maximum parsimony (MP), and Neighboring-Joining (NJ). The maximum likelihood (ML) method was far too sensitive to model specification (Goremykin et al. 2005), so we applied the BI method instead to analyze our data because it has a close connection to the ML method but with efficient computation (Rannala and Yang 1996). BI incorporates the Markov chain Monte Carlo (MCMC) process whose posterior probabilities are more reliable in measuring the accuracy of the estimated phylogeny (Rannala and Yang 1996). Moreover, BI has recently been introduced as a powerful method for reconstructing phylogeny (Holder and Lewis 2003).

The robustness of the tree nodes was evaluated with 1,000 bootstrap replicates in both MP and NJ analyses. A DNA substitution model for all codon positions and the first 2 codon positions of our data set was initially selected by use of Modeltest Version 3.7 (Posada and Crandall 1998) and the Akaike information criterion. Among the 56 models tested, the model of general time reversible (GTR), including rate variation among sites (+G) and invariable sites (+Γ) (=GTR + G + Γ), was chosen as the best fit for our 2 DNA data sets. We used MrBayes Version 3.1 (Ronquist and Huelsenbeck 2003) with MCMC process. The MCMC chains were started from a random tree and ran for 60,000 generations. Trees were sampled every 100 generations, and a consensus tree was built from all trees, excluding the first 100 (burn-in).

The MP tree was reconstructed using PAUP 4.0 (Swofford 2003) with the following options, a heuristic search and simple sequence addition; a tree bisection and reconnection branch-swapping algorithm, and a ''MulTrees'' option, in effect. The NJ trees were generated with distance matrices estimated by a K2P implemented in MEGA3 (Kumar et al. 2004).

Results and Discussion

Evolution of cpDNA Organizations in Seed Plants

Characteristics of C. taitungensis cpDNA

The complete cpDNA of *C. taitungensis* is a circular molecule of 163,403 bp (fig. 1) with the typical structures found in green algae and land plants: a pair of rRNA-containing IRs (IR$_A$ and IR$_B$; 25,074 bp each) separated by large single-copy (LSC; 90,216 bp) and small single-copy (SSC; 23,039 bp) segments. The cpDNA encodes 169 genes, including 123 protein-coding genes, 8 ribosomal RNA genes, and 38 tRNA genes. Twenty of them, including 12 protein-coding genes and 8 tRNA genes, contain introns (supplementary table 1, Supplementary Material online). The proportions of protein-coding region, rRNA, tRNA, intron, and intergenic spaces are 55.7%, 5.6%, 1.8%, 10.3%, and 26.6%, respectively. A total of 9 repeats, including 7 tandem repeats and 2 pairs of simple repeats (viz. the repeat 4 and 7), were found in the cpDNA of *C. taitungensis* (fig. 1; supplementary table 2, Supplementary Material online). Two of them (viz. the repeats 6 and 9) are located within the coding regions of genes. The repeat 6, residing at the 3′ end of *psbB*, are able to form a hairpin structure, suggesting that it might have a function for stabilization of mRNA structure. However, homologs of repeat 6 are exclusively found in other species of the *Cycas*. Overall, only the repeat 4 has highly similar sequences detected in the known cpDNAs of seed plants. Further investigation of this repeat might help to understand the evolution of repeat in cpDNA.

In addition, 36 open reading frames (ORFs) with a threshold of 120 bp are identified. Two of the ORFs, *orf107* and *orf75*, are reported for the first time. The former resides in the complementary strand of the *rrn23* gene and its amino acid sequence is 79%, which is similar to the

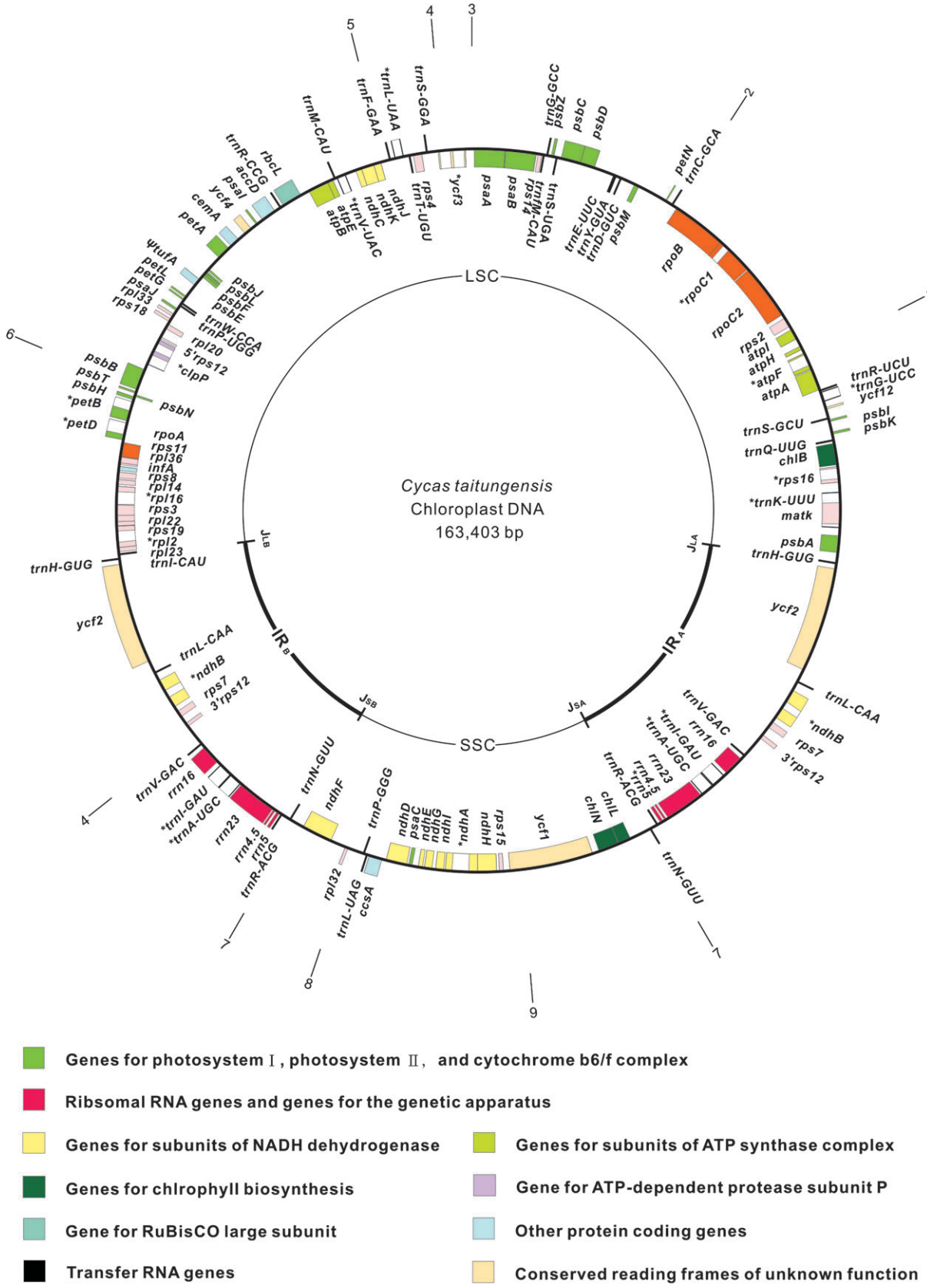

FIG. 1.—Gene map of *Cycas taitungensis* chloroplast genome. Genes shown on the inside and outside of the large circle are transcribed clockwise and counterclockwise, respectively. Asterisks denote spilt genes.

cp *orf91* of *Phalaenopsis aphrodite* (NC_007499.1), whereas *orf75* is located in the complementary strand of the *rrn16* gene and its predicted amino acid sequence is 97%, which is similar to an unknown protein (AAV44205) encoded in chromosome 5 of *Oryza sativa*. Transcripts of the *orf75* were experimentally verified in the *Cycas* chloroplast (supplementary fig. 2, Supplementary Material online). Because of the highly conserved nature of *rrn16* genes in all known land plants, homologues of *orf75* should be present in the corresponding genomic regions of all land-plant cpDNAs. However, the function of this ORF requires further study.

The cpDNA molecule of *C. taitungensis* is about 40 kb larger than those of the available 2 gymnosperms *Pinus thunbergii* and *Pinus koraiensis*, mainly because of the loss of all ndh genes and 1 IR region in the 2 pines. The gene number of *Cycas* (169 genes) is more than ferns (*Psilotum*, 150 genes; *Adiantum*, 131 genes) but fewer than that of pines (*P. thunbergii*, 201 genes; *P. koraiensis*, 205 genes). These data indicate that despite cp gene losses (largely transferred to the nucleus) during the evolution from green algae to ferns (Martin et al. 1998, 2002), among gymnosperms the number of protein-coding and tRNA genes appear to have increased in the conifer lineage.

In cpDNA organization, *C. taitungensis* is similar to *Ginkgo*, *Gnetum* (Wu CH, Wang YN, Chaw SM, unpublished data), and angiosperms in having a pair of IRs. Supplementary figure 1 (Supplementary Material online) shows a simplified phylogenetic tree of seed plants with a consensus topology of the 4 trees shown in figure 3, and the loss/retention/gain of protein-coding genes was mapped onto each cpDNA lineage. Note that *Cycas* and *Gingko* differ from 2 other gymnosperm lineages in their common retention of the Ψ*tufA* sequence. The conifer lineage (represented by *Pinus*) and *Gnetales* (represented by *Gnetum*) have concurrently lost all of the 9 *ndh* and *rps16* genes. The latter gene was also lost in parallel from *Psilotum*. Strikingly, *AccD*, *chlB*, *chlL*, *chlN*, *clpP*, *rps15*, and *rpl23*, which are present in the cycads, *Ginkgo*, and pine family, likely have been transferred to the nrDNA in *Gnetum*. Interestingly, the loss of *chlB*, *chlL*, and *chlN* genes also independently occurred in the *Psilotum* and *angiosperm* lineages. Complete cpDNA data from *Ginkgo*, non-Pinaceae conifers, and other Gnetophytes will fill in the gaps in our understanding of gene loss/retention during the seed plant evolution and their significance.

Synteny of an Ancient TufA Sequence in the cpDNAs of Cycads, Ginkgo, and a Hornwort

A 723-bp fragment, situated between the genic spacer of *psbE* and *petL* (fig. 1), was found to have 41% similarity to a protein synthesis elongation factor *tufA* (NC_008097.1) encoded in the cpDNA of *Chara vulgaris* (Turmel et al. 2006). Our PCR assays, involving a specific primer pair based on the conserved sequences of *psbE* and *petL*, further revealed homologs of this *tufA*-like gene in the cpDNAs of *Ginkgo* (accession number AB284317) and other 2 cycad families, Boweniaceae (represented by *Bowenia serrulata*; accession number AB284319) and Zamiaceae (represented by *Microcycas calocoma*, accession number AB284318). A Blast search of the predicted amino acid sequence of this 723 bp against each of the known, approximately 40 land-plant cpDNAs, revealed a homolog also present in the *Anthoceros formosae* (NCBI accession number 004543).

Compared with the sequences of available functional *tufA* genes (The alignment file is available in http://biodiv.sinica.edu.tw/research.php?pi=9), the *tufA*-like sequences in the cpDNAs of cycads, *Ginkgo*, and *Anthoceros* lack both start and stop codons and contain several stop codons within reading frames. Moreover, RT-PCR assay did not detect the presence of any *tufA* transcript (data not shown) in the cpDNAs of *Cycas* and *Ginkgo*, which implies that these *tufA*-like sequences are nonfunctional.

The functional *tufA* gene is encoded in the nrDNA of cyanobacteria and the cpDNAs of algae but is known to have been completely transferred from the cpDNAs to the nrDNA in rice, tobacco, legume family, and *Arabidopsis* (Baldauf and Palmer 1990; Sugita et al. 1994). Therefore, the nonfunctional *tufA* sequences found in the cpDNAs of *Anthoceros*, cycads, and *Ginkgo* are *tufA* pseudogenes (i.e., Ψ*tufA*). A phylogenetic tree reconstructed with the available functional *tufA* genes and our new data (fig. 2) uncovered that these 3 Ψ*tufA* sequences form a monophyletic group, are embedded within a clade composed of the cp *tufA* genes of green algae, and are most closely related to the cp *tufA* of *Chara*. This relationship suggests that the syntenic Ψ*tufA* genes in cpDNAs of *Anthoceros*, cycads, and *Ginkgo* are remnants of once-functional cp *tufA* genes and that they were descended from the charophyte lineage. It would be interesting to clone and verify the functional *tufA* genes in the nucleus of cycads and *Ginkgo* to see whether the transfer of *tufA* genes in the 2 lineages is independent from that of other seed plants.

Notably, gene arrangement of *psbE*, Ψ*tufA*, *petL*, and *petG* is exactly the same in the cpDNAs of *Anthoceros*, *Cycas*, and *Ginkgo*. Excluding Ψ*tufA*, this gene order is also highly conserved in the cpDNA of land plants. But in *Chara vulgaris*, the gene order of *tufA*, *petL*, and *petG* is inverted. Therefore, an inversion event likely occurred either in the land plants or charophytes after the 2 lineages diverged from each other in the Silurian era, approximately ~430–408 MYA. Further comparative study of cpDNA organization of green algae and cyanobacteria are required to disclose this inversion scenario.

tRNAPro (GGG) Gene Was Lost from the Angiosperm Lineage

The *trnP-GGG* was first found residing between the *trnL* and *rpl32* genes of *P. thunbergii* (Wakasugi et al. 1994). Recently, it was also annotated in the corresponding genomic locations of some charophytes (Turmel et al. 2002a, 2002b, 2006), *Physcomitrella patens* (a moss; but as a pseudogene), *Huperzia lucidula* (a lycopod), *Psilotum nudum*, and *Adiantum capillus* (ferns). Here, we report its presence in the cpDNAs of *Cycas*, *Gnetum*, and *Ginkgo* (AB295960). Therefore, the location of *trnP-GGG* was syntenic from green algae to primitive land plants until the emergence of seed plants, but the gene was subsequently lost from the angiosperm lineage at least 150 MYA (Chaw et al. 2004).

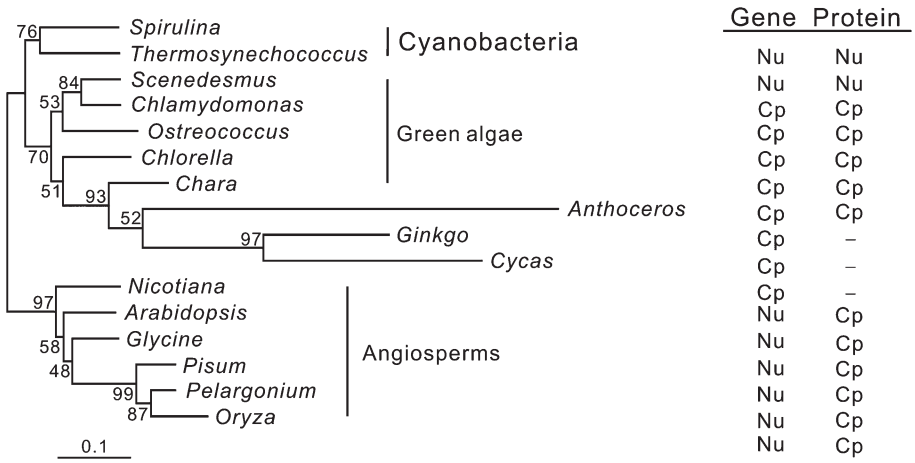

FIG. 2.—A NJ tree reconstructed with K2P model (Kimura 1980) with use of all codon positions of *tufA* genes from 2 cyanobacteria (*Thermosynechococcus* and *Spirulina*), 4 algae (*Scenedesmus*, *Chlamydomonas*, *Ostreococcus*, and *Chlorella*), one hornwort (*Anthoceros*), 2 gymnosperms (*Cycas* and *Ginkgo*), and 6 angiosperms (*Nicotiana*, *Pisum*, *Arabidopsis*, *Glycine*, *Pelargonium*, and *Oryza*). The right table shows the location of each *tufA* gene and its protein. Note that in the tree, the Ψ*tufA* genes of *Anthoceros*, *Cycas* and *Ginkgo* are not sister to the functional *tufA* genes of angiosperms but to the *Chara* instead.

Only One RNA Editing Site Predicted

Among the 169 protein-coding genes/ORFs, only the second nucleotide of the *petL* gene is predicted to be altered by RNA editing, resulting in a C-to-U change and resurrection of the start codon. This editing site appears to be a common character in the *Pinus* (Wakasugi et al. 1996), *Cycas*, and *Ginkgo* (DQ069698) lineages. Compared with the number of cpDNA editing sites in *P. thunbergii* (26 sites, Wakasugi et al. 1996) and other known land plants, the number of cpDNA editing sites in *Cycas* (only one site) is strikingly reduced. It is interesting to note that quite a high frequency of RNA editing had been reported for *Cycas revoluta* in mitochondrial sequences (Malek et al. 1996), whereas, proportionality of RNA editing frequency between chloroplast and mitochondrial transcripts is generally observed in plant taxa. If the plant chloroplast RNA editing is monophyletic in origin, as suggested by Tillich et al. (2006), the great reduction in number of editing sites in *Cycas* may indicate that editing is no longer needed in this lineage. More cpDNA data from other cycads are critically needed to verify whether this reduction in RNA editing sites is a common state in the cycad lineage or specific only to *Cycas*.

cpDNA Phylogeny Suggests that Extant Gymnosperms and Angiosperms Are Separate Monophyletic Clades

After removal of all gaps, unknown sites, start and stop codons, and positions with difficulty to align, 30,273 bp were used for comparison and reconstruction of phylogenetic trees. The variable sites were 16,314 bp, of which 12,023 were parsimony informative. In the phylogenetic analyses, 2 bryophytes, *Marchantia polymorpha* and *Physcomitrella patens* (supplementary table 3, Supplementary Material online), were used as the outgroups, and a lycophyte, *Huperzia lucidula*, and a fern, *Psilotum nudum* (supplementary table 3, Supplementary Material online), as the reference taxa. The 4 phylogenetic trees (fig. 3*A*–*D*) depict the 2 BI trees based on all codon positions (BI-all) and the first 2 codon positions (BI-1 + 2), the MP tree (MP), and the NJ tree based on Tv sites (NJ-Tv) of the first 2 codon positions and the K2P model. Both BI trees (figs. 3*A* and *B*) used a GTR + G + Γ DNA substitution model, which was selected as the best model by using Modeltest (Posada and Crandall 1998).

Because the third codon positions of cp protein-coding genes of the sampled taxa were saturated with substitutions (data not shown) and have exceptionally accelerated substitution rate in *Gnetum* (table 1; see section Substitution Rates and Long Branch Attraction. . . below), they were excluded from the reconstruction of BI-1 + 2, MP, NJ-Tv, and NJ-Tv + Ts trees. In the NJ-Tv tree, only the Tv sites were used because the Ts sites of *Gnetum* and several seed plant lineages (e.g., *Pinus*, *Gnetum*, *Oryza*, and *Phalaenopsis*) contain conflicting phylogenetic signals (table 1; see also Chang et al. 2006).

Topologies of the 2 BI trees and the NJ-Tv trees (fig. 3*A*, *B*, *D*) are nearly identical except at the deepest divergence of angiosperms—the relative positions of *Amborella* and *Nymphaea* clades—and the diversification within rosids. In addition, the bootstrap supports for the monophyly of gymnosperms are moderate in the NJ-Tv tree. Agreements among the 4 trees in figure 3 include 1) the extant seed plants forming a monophyletic clade; 2) a sister relationship between the extant gymnosperms and angiosperms; 3) robust bootstrap supports for the *Cycas*–*Ginkgo* clade, the *Pinus*–*Gnetum* clade, and several major angiosperm clades, including the monocots, the eudicots, the rosids; and 4) relationships within main groups of angiosperms being highly supported at most nodes. The MP analysis yielded single tree (fig. 3*C*) of 18,093 steps. The branching pattern of the MP tree differs from that of other 3 trees in the relative positions of the lycophyte (represented by *Huperzia*) and the fern (represented by *Psilotum*) lineages, and the placement of a primitive magnoliid (the *Calycanthus*). Nevertheless, topology of this MP tree appears to be unreliable because ferns are unlikely more basal than

Table 1
Tajima's Relative Rate Tests between *Cycas*, *Gnetum*, and Other 5 Seed Plant Lineages Based on Concatenated 56 cp Protein–Coding Genes

Taxa Compared		Reference Taxon	P of χ^2 Test								
1	2	3	1st-Ts	1st-Tv	1st-Ts + Tv	2nd-Ts	2nd-Tv	2nd-Ts + Tv	3rd-Ts	3rd-Tv	3rd-Ts + Tv
Cycas	*Ginkgo*	*Psilotum*	0.926	0.774	0.967	0.067	0.611	0.058	0.262	0.448	0.430
Cycas	*Gnetum*	*Psilotum*	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.047[a]	0.000[a]	0.000[a]
Cycas	*Pinus_t*	*Psilotum*	0.022[a]	0.000[a]	0.000[a]	0.011[a]	0.000[a]	0.000[a]	0.500	0.000[a]	0.034[a]
Cycas	*Amborella*	*Psilotum*	0.001[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]
Cycas	*Oryza*	*Psilotum*	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]
Cycas	*Phalaenopsis*	*Psilotum*	0.003[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.001[a]	0.000[a]	0.000[a]
Cycas	*Nicotiana*	*Psilotum*	0.016[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.000[a]	0.018[a]	0.000[a]	0.000[a]
Gnetum	*Pinus_t*	*Psilotum*	0.000[a]	0.000[a]	0.000[a]	0.053	0.000[a]	0.000[a]	0.023[a]	0.000[a]	0.000[a]
Gnetum	*Amborella*	*Psilotum*	0.012[a]	0.000[a]	0.000[a]	0.845	0.000[a]	0.000[a]	0.500	0.000[a]	0.000[a]
Gnetum	*Oryza*	*Psilotum*	0.593	0.000[a]	0.000[a]	0.208	0.000[a]	0.013[a]	0.088	0.000[a]	0.025[a]
Gnetum	*Phalaenopsis*	*Psilotum*	0.006[a]	0.000[a]	0.000[a]	0.874	0.000[a]	0.000[a]	0.644	0.000[a]	0.000[a]
Gnetum	*Nicotiana*	*Psilotum*	0.003[a]	0.000[a]	0.000[a]	0.519	0.000[a]	0.000[a]	0.537	0.000[a]	0.000[a]

[a] χ^2 test significant at the 5% level (i.e., $P < 0.05$); *Pinus_t*: *Pinus thunbergii*

lycophytes (see also Raubeson and Jansen 1992b; Stewart and Rothwell 1993). The NJ-Tv + Ts tree based on the Tv and Ts sites of the first 2 codon positions is shown in supplementary figure 4 (Supplementary Material online). In this tree, bootstrap supports for the seed plants as a whole, the basal-most dicots, and the rosids are low. Additionally, topology of this tree highly resembles that of the MP tree, especially in the relative positions of *Huperzia* and *Psilotum*, which implies that this NJ-Tv + Ts tree is also untrustworthy.

Our analyses of cpDNA sequence data strongly indicate that the extant seed plant lineages are monophyletic, with the ferns being a sister group. In other words, living seed plants have a common origin and the seed evolved only once. This result is consistent with molecular analyses of single (Hasebe et al. 1992; Goremykin et al. 1996; Chaw et al. 1997) or multigenes (Bowe et al. 2000; Chaw et al. 2000) and many subsequent studies (reviewed in Burleigh and Mathews 2004). However, it contradicts the early analyses of Hamby and Zimmer (1992) and their subsequent authors such as Albert et al. (1994), Rydin et al. (2002), and Rai et al. (2003), in which Gnetales was suggested to be sister to other seed plants, rendering the extant seed plants paraphyletic. Because of space limitation, we defer to Chaw et al. (1997), Burleigh and Mathews (2004), and Hajibabaei et al. (2006) for a discussion on the competing hypotheses for the evolutionary relationships of 5 major seed plant lineages.

The Deepest Split in the Evolution of Extant Gymnosperms Is between the *Cycas–Ginkgo* and the Gnetales–Pinales Clades

Figure 3 shows that the 4 trees based on 3 methods consistently resolved *Cycas–Ginkgo* and *Gnetum–Pinus* as 2 separate monophyletic clades, which indicates that these 2 clades represent the deepest split in the evolutionary diversification of extant gymnosperms. As well, the diversification of modern gymnosperm lineages appear to be rather fast or involve fewer nucleotide changes than in the branch leading to the angiosperms. This phylogeny and the short branch length before the diversification node of gymnosperms are also supported by the phylogenetic analyses of chloroplast intergenic transcribed spacer data (Goremykin et al. 1996) and 18S rRNA data (Chaw et al. 1997). Because of space limitations, we defer to Chaw et al. (1997) for a discussion of reproductive resemblance between cycads and *Ginkgo*. However, our results conflict with those from more recent multigene analyses of Bowe et al. (2000), Chaw et al. (2000), Nickrent et al. (2000), and Soltis et al. (2002), among many others, in which cycads were suggested to be the basal-most clade in the gymnosperm phylogeny, followed by *Ginkgo*.

Substitution Rates and Long Branch Attraction

Nearly Equal Rates in Cycas and Ginkgo

The branch lengths shown in figure 3, clearly reveal considerable variations in substitution rates among seed plant lineages. The rates in the 4 gymnosperm lineages are in the order of *Cycas*/*Ginkgo* < pines < *Gnetum*. In table 1, *Psilotum* was used as the reference taxon to examine the relative substitution rates between *Cycas* and *Ginkgo*, *Gnetum*, *Pinus*, *Amborella* (basal-most dicots), *Oryza* (a fast evolving monocot), *Phalaenopsis* (a medium evolving monocot), and *Nicotiana* (an eudicot with medium evolving rate), respectively, as well as between *Gnetum* and the latter 5 lineages mentioned above, separately. Table 1 shows that in all codon positions, the substitution rates of *Cycas* and *Ginkgo* do not evolve differently (all $P > 0.05$) and are the slowest among the seed plants except for at the Ts sites of the 3rd codon positions, which are not significantly slower than in *Pinus*.

Accelerated Rates in Gnetum and LBA

In contrast, the Tv sites of all codon positions have the fastest rates in the *Gnetum* lineage. Of greatest note, *Gnetum*'s Ts rates apparently are not parallel with its accelerated Tv rates. Specifically, *Gnetum*'s Ts rates for the 1st codon positions are not significantly higher than those of the *Oryza* lineage. Similarly, *Gnetum*'s Ts rates do not differ from those of *Pinus*, *Amborella*, *Oryza*, *Phalaenopsis*, and *Nicotiana* in the 2nd codon positions and do not differ from rates in the latter 4 taxa in 3rd codon positions.

The above data might explain why the topology of NJ-Tv + Ts tree (based on the first 2 codon positions; supplementary fig. 4, Supplementary Material online) deviates

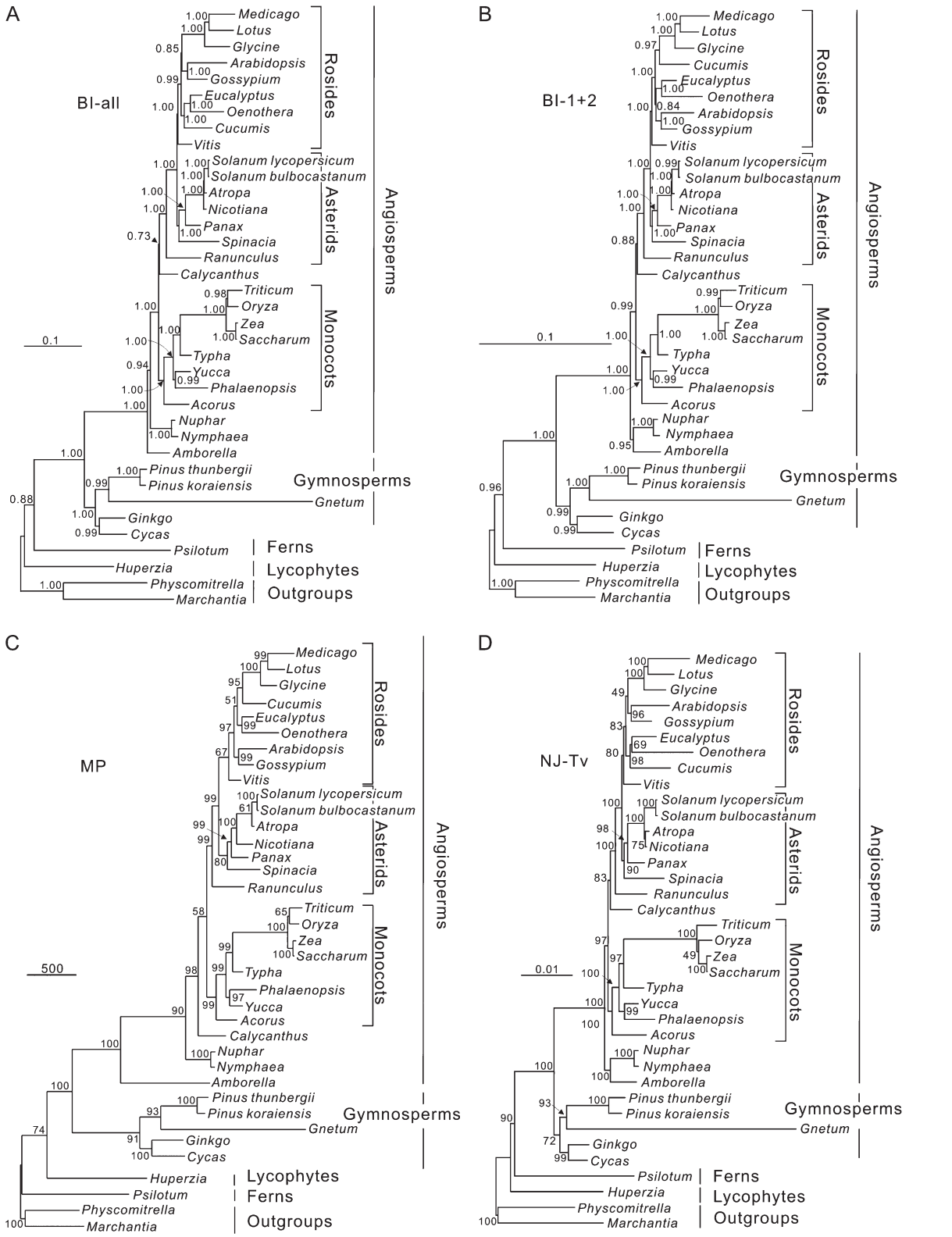

FIG. 3.—Phylogenies of 37 land-plant cpDNAs based on concatenated, 56-chloroplast protein-coding genes (supplementary table 4, Supplementary Material online). *Physcomitrella* and *Marchantia* were used as the outgroups. (*A*) The best BI tree inferred from all 3 codon positions. Trees of (*B*)–(*D*) were reconstructed with use of the 1st and 2nd codon positions. (*B*) The best BI tree. (*C*) The single most parsimonious tree. It has a length of 18,093 steps, consistency index (CI) = 0.553, retention index (RI) = 0.657, and homoplasy index (HI) = 0.047. (*D*) The NJ tree inferred from the transversional sites.

from the NJ-Tv tree in figure 3*D*. In other words, in the NJ-Tv + Ts tree the inclusion of Ts sites from the first 2 positions has rendered the *Gnetum* a sister of the other seed plants. Conflicting evolutionary rates observed at the Ts and Tv sites of the first 2 codon positions in our large data set are consistent with the findings of Chang et al. (2006), who analyzed 20 cpDNA species of seed plants and sampled only one gymnosperm, the black pine.

Unequal substitution rates of sequences are well known to affect tree reconstruction algorithms, and LBA (Felsenstein 1978) can lead to a false clustering of the longest branches, regardless of the underlying phylogeny (Felsenstein 1978; Bergsten 2005). Nonetheless, we are convinced that LBA might not influence our data set. As reported earlier in the plastid genomes (Whitfeld and Bottemley 1983; Chaw et al. 2004), a bias in the nucleotide composition is also detected at the 3rd codon positions (A = 32%, C = 14%, G = 16%, T = 39%; χ^2 test, $P < 0.000$) of the presently studied seed plants. However, the positions have been excluded from the reconstruction of one of the BI trees and both MP and NJ trees (fig. 3). Moreover, because of the general concordance of our analyses based on 3 methods with high bootstrap supports and their consistence with phylogenies of other molecular data (e.g., *rbcL* [Hasebe et al. 1992], cpITS sequences [Goremykin et al. 1996], nuclear 18S rRNA sequences [Chaw et al. 2000]), and cp structural changes mentioned previously, our tree topologies (fig. 3) are unlikely to be affected by LBA. The moderately supported monophyly of gymnosperms shown in figure 3*D* is likely due to the fewer sites (i.e., only the Tv sites) used in the tree reconstruction.

Why *Gnetum* (or Gnetales) has such a high evolutionary rate has not been investigated. The genus consists of about 33 species; most are large woody climbers and 2 are trees. There are 10 species occur in tropical South America, 1 in West Africa, and the remainder in Tropical and subtropical Asia (Won and Renner 2006). In general, the habits and habitats of the genus differ immensely from most other gymnosperms except for some subtropical and tropical Podocarpaceae (such as *Dacrycarpus*). The *Gnetum* species have to compete for light with species of the crown layer in subtropical and tropical forests, which are usually dominated by angiosperms. Our field and greenhouse observations indicate that for *G. parvifolium*, for example, only 6–7 years are required to produce male and female cones after seed germination, and adult plants set mature cones yearly. On the contrary, for cycads, *Ginkgo*, and conifers, at least 10 or even 30 years are required for development to reach reproductive stages and cones require at least 1–3 years to set mature seeds. We suspect that the comparatively short generation time and faster seed production of *Gnetum* might have an effect on its accelerated rates. However, which gene or gene groups of its cpDNA have been influenced by the generation time and/or environmental selection force needs further investigation. Furthermore, Tv mutation of the first 2 codon positions and Ts of the 2nd position in a protein-coding gene will cause amino acid changes. How *Gnetum*'s cpDNA genes underwent at such a fast mutation rate and what selection forces triggered the massive mutations remain enigmatic and intriguing.

Hoegg et al. (2004) suspected that "tree reconstruction algorithms tend to regularly assign shorter branches to basal taxa when these are placed paraphyletically toward a taxon-rich and well-supported crown group." Coincidently, this situation is also apparent from the short branches of the *Cycas–Ginkgo* clade, the 2 reference taxa (*Huperzia* and *Psilotum*), and the outgroups in our 3 trees (fig. 3*B-D*). Hence, we echo the authors' remarks that "the potential impact of this phenomenon on tree-based molecular methods appears to be an interesting problem to be addressed in future studies."

Comparative Structural Changes of cpDNAs Support the cpDNA Phylogeny

Gene Orders near IR/LSC Junctions Contain Useful Phylogenetic Information

The locations of IR and LSC junctions have been known to vary among groups of angiosperms (Maier et al. 1995; Goulding et al. 1996; Perry et al. 2002; Kim and Lee 2004; Chang et al. 2006). Figure 4 details a comparison of the locations of IR–LSC junctions (i.e., JLA and JLB) and their flanking genes among major extant lineages of land plants, including ferns (*Psilotum*), gymnosperms (*Cycas*, *Ginkgo*, *Gnetum*, and *Pinus*) and 3 angiosperm representatives—*Amborella* (a basal-most dicot), *Nicotiana* (a eudicot), *Oryza* (Poaceae), and *Phalaenopsis* (a genus of Orchidaceae). With *Psilotum* used as an outgroup, monophyly of the seed plants is congruent with a unique duplication of *ycf2* and shift of this gene from the LSC region to the IR regions (see also the section Duplication of Ycf2 Gene in IRB Regions Predates the Divergence of Seed Plants Rather Than Leafy Plant). Furthermore, monophyly of the 4 gymnosperm orders is entirely concordant with their common possession of a duplicate *trnH* gene located exactly in the upstream of *ycf2*, although in the *Pinus* these 2 genes have been lost from the IRB regions. Figure 4 also shows that in contrast to the IRs of ferns and gymnosperms, those of the 3 angiosperm representatives have further expanded with conversion of more LSC genes (e.g., *rpl2* and *rpl23*) that reside near the IR–LSC junctions.

Indels and Gene Loss/Retention Lend Evidence to the Phylogeny within Gymnosperms

Within the modern gymnosperms, a sisterhood relationship between *Cycas* and *Ginkgo* is sustained by their shared 10 unique genomic characters—9 unique indels (5 insertions of 3–6 nt and 4 deletions of 3–9 nt) in the 56 protein-coding gene data set. Mapping the states of loss/retention/gain of the cp protein–coding genes onto a simplified consensus tree of figure 3*A*, *B*, and *D* revealed several unambiguous patterns that support the monophylies of seed plants and angiosperms, as well as relationships within 2 clades of gymnosperm orders (*Gnetum–Pinus* and *Cycas–Ginkgo*). As shown in Supplementary figure 1 (Supplementary Material online), the monophyly of *Cycas* and *Ginkgo* is evidenced by the synteny of a *ΨtufA* sequence (fig. 2) that descended from the cp *tufA* gene of green algae. However, a sister relationship between *Gnetum* (Wu CH, Wang YN, Chaw SM, unpublished data) and *Pinus* is well supported by their concurrent loss of *rps16* and all of the 9

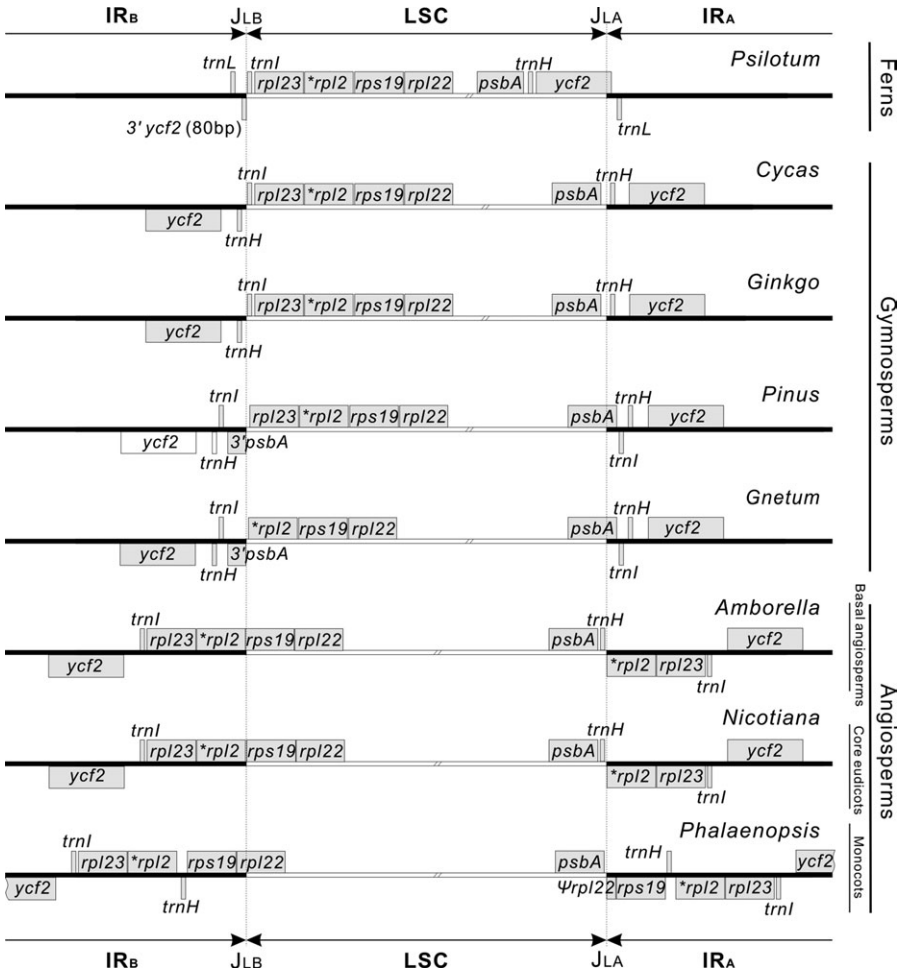

FIG. 4.—Comparison of the genes flanking the IR–LSC junctions (JLA and JLB) among *Psilotum* (a fern; as the outgroup to seed plants), *Cycas*, *Ginkgo*, *Pinus*, *Gnetum* (gymnosperms), *Amborella* (a basal angiosperm), *Nicotiana* (a eudicot), and *Phalaenopsis* (a monocot). Gray boxes denote gene names and their presence; a blank box denotes absence of the designated gene and its likely ancestral location at the IR.

ndh genes and by their common expansion of IRs, which also encompass a *trnI-UUG* gene and a partial 3′ *psbA* sequence (fig. 4).

The unusual occurrence and position of the *trnI*-3′ *psbA* sequence cluster near the IR/LSC junction of *Pinus* were first discovered by Tsudzuki et al. (1992). Our preliminary survey indicates that IRs of *Welwitchia*, *Picea* (a genus of Pinaceae), and the other conifer genus, *Taiwania* (Wu CH, Wang YN, Chaw SM, unpublished data), also contain such characteristics. These findings suggest that the origin of this gene cluster probably predates the divergence of Gnetales and Pinales. Of note, in *Psilotum*, *trnI* and *ycf2* are situated in LSC and transcribed in the same orientation, whereas in the seed plants they are either near JLB or in IRB and transcribed oppositely. These data suggest that before the diversification of seed plants, a conversion event had taken place near the IR/LSC junctions.

Duplication of Ycf2 Gene in IRB Regions Predates the Divergence of Seed Plants Rather Than Leafy Plants

Each IR of *Cycas* cpDNA possesses a copy of *ycf2* similar to the known cpDNAs of angiosperms. Our PCR assay (supplementary fig. 3, Supplementary Material online) confirms that one copy of *ycf2* is also present in each IR of *Gnetum* and *Ginkgo*. In addition, *ycf2* is found in both IRs of a leptosporangiate fern, *Adiantum capillus–veneris*, but absent from IRs of primitive land plants such as liverworts (*Marchantia*), hornworts (*Anthocerus*), mosses (*Physcomitrella*), club mosses (*Huperzia*), and an eusporangiate fern (*Psilotum*). However, duplication and shift of *ycf2* to the IRs of *Adiantum* could be an independent case and only one copy of *ycf2* was likely present in the ancestral cpDNAs of ferns because 1) *Psilotum*, one of the basal-most ferns, possesses a pair of IRs without an intact *ycf2* but a short 3′ fragment of the gene (fig. 4), 2) the length of IR regions in *Osmunda* (a primitive fern) is reported to be only about 13 kb (Stein et al. 1992), which seems too short to include an intact *ycf2* gene (~7 kb) plus a common ribosomal RNA operon (~8 kb), and 3) the extant fern orders are robustly supported as a monophyletic lineage by multigene analyses (Pryer et al. 2001, 2004; Schneider et al. 2004).

An Alternative Model for the Loss of Large IR regions in the Pinus *cpDNA*

Strauss et al. (1988) first reported the lack of an IR but extensive rearrangement in the cpDNAs of 2 pines (*Pinus*

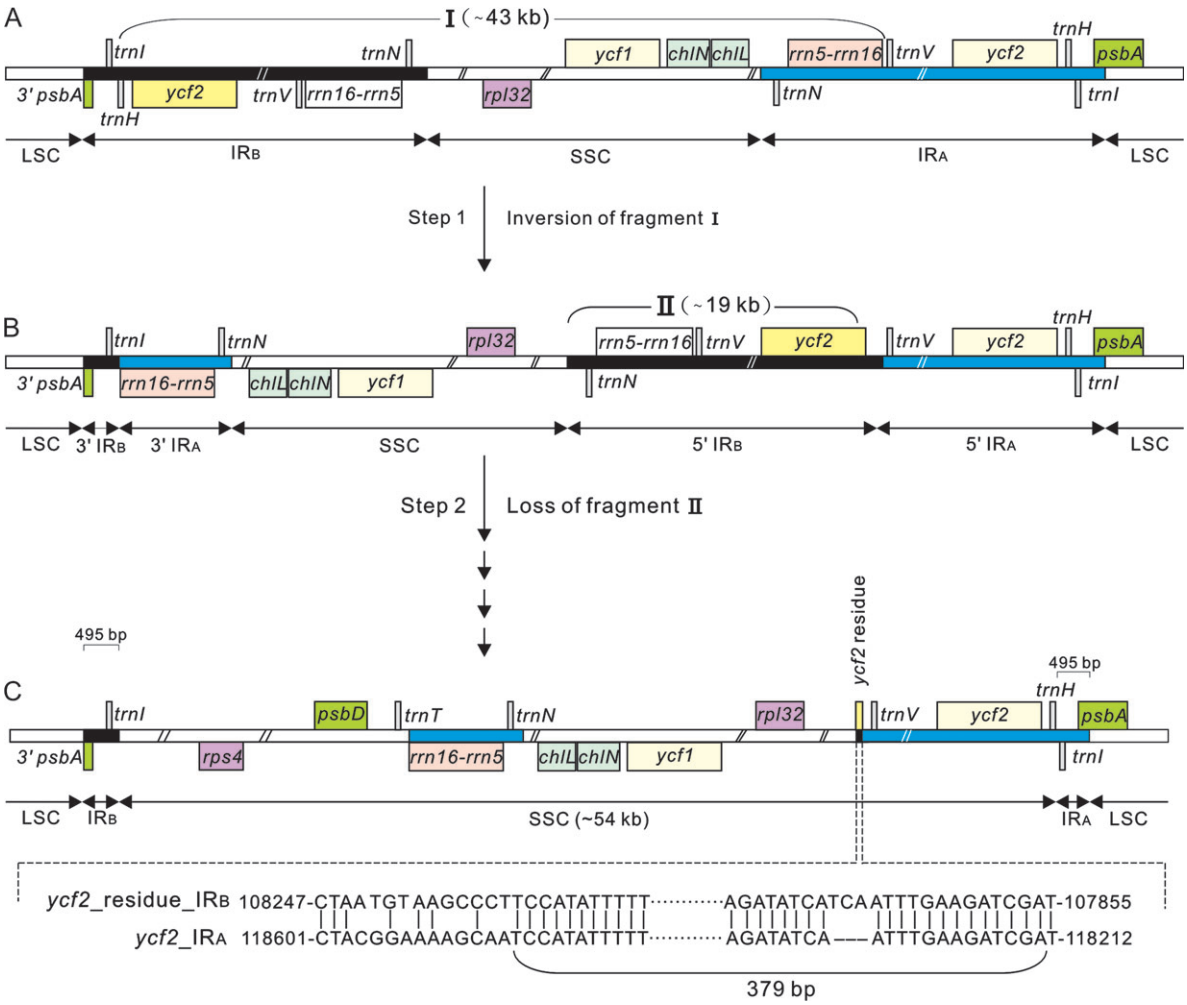

FIG. 5.—A 2-step model for the loss of the IRB region in the ancestral cpDNA of *Pinus thunbergii*. (*A*) The hypothesized ancestral form; (*B*) the Ts form; and (*C*) the present form. A sequence remnant (379-bp long; top row) of the ancestral *ycf2* originally belonging to the IRB is alignable with the functional *ycf2* sequence (bottom row) in the IRA. "|": identical base pairs; "-": a gap filled to perfect alignment.

radiata D. Don and *Pseudotsuga menziesii* [Mirb.] Franco). The authors considered that the loss of IR was proceeded by 2 steps: deletion of a portion of an IR segment and loss of the entire IR segment, resulting in one copy of the original IR intact but no repeat elsewhere in the genome. On the basis of cpDNA structural mapping, Raubeson and Jansen (1992a) also considered a shared loss of one large IR in all conifers, whereas Tsudzuki et al. (1992) proposed an evolutionary pathway to explain the retention of a 495-bp IR in the present cpDNA of *P. thunbergii* (black pine). The authors argued that the ancestral cpDNA of black pine "contained an IR encompassing a 3′ *psbA* and an rRNA gene cluster" and that one segment of IRB, from 495 bp (Tsudzuki et al. 1992) downstream of JLB, was deleted. Presumably, this incomplete deletion in black pine resulted in the IR regions being reduced to 495 bp and the lack of one rRNA gene cluster (fig. 5*C*).

A close inspection of the cpDNA structure of black pine with that of available reference species from *Chara*, *Physcomitrella*, *Huperzia*, *Cycas*, *Amborella*, and *Nymphaea* yields an alternative model to better interpret the scenarios resulting in the current cpDNA organization of black pine. Figure 5 shows a 2-step model to derive the presently observed cpDNA map of black pine (fig. 5*C*), whereby 2 rather short fragments (495 bp) retained from the ancestral IR regions are separated by an inverted SSC (relative to that in fig. 5*A*) of about 54 kb and a residue sequence (379 bp) of *ycf2* that presumably was situated in the ancestral IRB (fig. 5*B*). Note that this 379-bp residue is positioned between the *rpl32* of SSC and *trnV* of IRA.

Figure 5*A* illustrates the IRB, SSC, and IRA segments of a hypothetical ancestral cpDNA of black pine. Each IR contains a partial 3′ *psbA* sequence, a *trnI*, a *trnH*, and a *ycf2* bordering on LSC. During evolution, a transitional form likely existed (fig. 5*B*), in which the fragment I (fig. 5*A*) that covers nearly the whole region of ancestral IRB, ancestral SSC, and the *rRNA* operon and its upstream region of ancestral IRA was inverted. As a result, figure 5*B* shows the gene order of 3′IRB-SSC-5′IRB prominently different from that in figure 5*A*. The fragment I is estimated to be about 43-kb long. Subsequently, in step 2 the fragment II (fig. 5*B*; estimated to be ca. 19 kb), which comprises a large segment of inversed IRB, was lost, generating the extant cpDNA of black pine shown in figure 5*C*. Note that an

apparent IRB footprint—the 3′ end of *ycf2* residue (379 bp)—is alignable with the current *ycf2* (originally residing in IRA) sequence located in SSC.

The inversion event of fragment I is also clearly evident by 3 observations: 1) In the cpDNA of black pine, the rRNA operon and *ycf2* (fig. 5*C*) are arranged in the opposite orientation, whereas in all IRs of the reference cpDNAs, the rRNA operons and *ycf2* genes are in the same orientation; 2) In SSC regions of the reference species, gene orders of the cluster *rpl32*, *ycf1*, *chlL*, and *chlN* (fig. 5*A*) are inverted in the black pine (fig. 5*C*); and 3) In IRA of the reference species (as in fig. 5*A*), the rRNA operon that contains the gene clusters *rrn5*, *rrn4.5*, *rrn23*, and *rrn16* (simplified as *rrn5-rrn16*) is downstream of *trnN*. However, in the black pine, the rRNA operon is oriented opposite of *rrn16-rrn5* and positioned upstream of *trnN* (fig. 5*C*).

Conclusions

We determine and analyze the surprising structure of the cpDNA of the first cycad, *C. taitungensis*, and 56 cp protein–coding genes of a Gnetales representative, *G. parvifolium*. These analyses help to fill in the gap of information about gene content and organizations between fern and angiosperm cpDNAs. In the cpDNA organization, *Cycas* is more similar to ferns and *Ginkgo* than to pines in the presence of 2 large IR regions. An ancient signature of green algae's cp *tufA* sequences is first reported in the cpDNAs of cycads, *Ginkgo*, and a hornwort. However, both cpDNAs of pines and *Gnetum* resemble each other in the loss of all *ndh* genes and shift of IR–LSC junctions, which suggests a close relationship between the 2 lineages.

Our phylogenetic analyses of a larger cpDNA data set (37 taxa) further strengthen the view that extant gymnosperms constitute a monophyletic clade and that modern seed plants share a common ancestor and the seed evolved only once (Rothwell 1981, 1982; Chaw et al. 1997). In addition, an alternative model for the loss of 2 large IR regions in the *Pinus* is proposed. In good agreement with the study of Raubeson and Jansen (2005), we further demonstrate that comparative cpDNA organization, specifically the gene orders bordering IR–LSC junctions, is a very powerful tool for reconstructing ancient evolutionary relationships in seed plants.

Although we cannot justify the gnetifer and gnepine hypotheses in this study, more cpDNA data from non-Pinaceae conifers can help us settle the issue. Our lab is currently expanding the survey of cpDNAs across the diversified gymnosperms for a better understanding of the cpDNA evolution and for readdressing old questions in seed plant phylogeny. This study also attests to the utility of concatenated protein-coding sequences in addressing the deep phylogeny of 5 main seed plant lineages and the gymnosperm orders. However, when the NJ method is applied, codon positions of cp protein–coding genes should be used with caution because the Ts and Tv sites at the first 2 codon positions contain conflicting signals and their 3rd codon positions are biased with nucleotide compositions. In addition substitution rates are highly variable among seed plants.

Supplementary Material

Supplementary tables 1–4 and figures 1–4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a research grant from Research Center for Biodiversity, Academia Sinica to S.M.C. and in part by a grant from National Science Council to Y.N.W. We thank the administrator of South China Botanical Garden for the kind gifts of plant materials. This research is in partial fulfillment of the PhD program in the School of Forestry and Resource Conservation, National Taiwan University for C.S.W. Special thanks are due to the 2 anonymous reviewers who provided critical and helpful suggestions to the authors.

Literature Cited

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. Nature. 387:489–493.

Albert VA, Backlund A, Bremer K, Chase MW, Manhart JR, Mishler BD, Nixon KC. 1994. Functional constraints and *rbcL* evidence for land plant phylogeny. Ann Mo Bot Gard. 81:534–567.

Baldauf SL, Palmer JD. 1990. Evolutionary transfer of the chloroplast *tufA* gene to the nucleus. Nature. 15:262–265.

Bergsten J. 2005. A review of long-branch attraction. Cladistic. 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 USA. 97:4092–4097.

Brenner ED, Stevenson DW, Twigg RW. 2003. Cycads: evolutionary innovations and the role of plant-derived neurotoxins. Trends in 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.

Chang CC, Lin HC, Lin IP, et al. (11 co-authors). 2006. The chloroplast genome of *Phalaenopsis Aphrodite* (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Mol Biol Evol. 23:279–291.

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. 58:1–18.

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 USA. 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.

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

Gallois JL, Achard P, Green G, Mache R. 2001. The Arabidopsis chloroplast ribosomal protein L21 is encoded by a nuclear gene of mitochondrial origin. Gene. 274:179–185.

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

Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol. 22:1813–1822.

Goulding SE, Olmstead RG, Morden CW, Wolfe KH. 1996. Ebb and flow of the chloroplast inverted repeat. Mol Gen Genet. 252:195–206.

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.

Hamby RK, Zimmer EA. 1992. Ribosomal RNA as a phylogenetic tool in plant systematics. In: Soltis PS, Soltis DE, Doyle JJ, editors. Molecular systematics of plants. New York: Chapman and Hall. p. 50–91.

Hasebe M, Ito M, Kofuji R, Iwatsuki K, Ueda K. 1992. Phylogeny of gymnosperms inferred from *rbcL* gene sequences. Bot Mag Tokyo. 105:673–679.

Hill KD, Chase MW, Stevenson DW, Hill HG, Schutzman B. 2003. The families and genera of cycads: a molecular phylogenetic analysis of cycadophyta based on nuclear and plastid DNA sequences. Int J Plant Sci. 164:933–948.

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

Hoegg S, Vences M, Brinkmann H, Meyer A. 2004. Phylogeny and comparative substitution rates of frogs inferred from sequences of three nuclear genes. Mol Biol Evol. 21:1188–1200.

Holder M, Lewis PO. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet. 4:275–284.

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

Kim KJ, Lee HL. 2004. Complete chloroplast genome sequences from Korean ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. 11:247–261.

Kimura M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 16:111–120.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol. 16:573–576.

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

Maier RM, Neckermann K, Igloi GL, Kossel H. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. J Mol Biol. 251:614–628.

Malek O, Lattig R, Hiesel K, Brennicke A, Knoop V. 1996. RNA editing in bryophytes and a molecular phylogeny of land plants. EMBO J. 15:1403–1411.

Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 2002. Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA. 99:12246–12251.

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.

Nicholas KB, Nicholas HB Jr. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Available at: http://www.nrbsc.org/gfx/genedoc/index.html. Accessed 2007 May 7.

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

Norstog KJ, Nicholls TJ. 1997. The biology of the cycads. Ithaca: Cornell University Press.

Perry AS, Brennan S, Murphy DJ, Kavanagh TA, Wolfe KH. 2002. Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. DNA Res. 31:157–162.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Pryer KM, Schneider H, Smith AR, Cranfill R, Wolf PG, Hunt JS, Sipes SD. 2001. Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature. 409:618–622.

Pryer KM, Schuettpelz E, Wolf PG, Schneider H, Smith AR, Cranfill R. 2004. Phylogeny and evolution of ferns (monilophytes) with a focus on the early leptosporangiate divergences. Am J Bot. 91:1582–1598.

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.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol. 43:304–311.

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

Raubeson LA, Jansen RK. 1992b. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science. 255:1697–1699.

Raubeson LA, Jansen RK. 2005. Chloroplast genomes of plants. In: Henry RI, editor. Plant diversity and evolution: genotypic and phenotypic variation in higher plants. Wallingford (UK): CABI. p. 45–68.

Robinson SP, Downton WJ. 1984. Potassium, sodium and chloride content of isolated intact chloroplasts in relation to ionic compartmentation in leaves. Arch Biochem Biophys. 228:197–206.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rothwell GW. 1981. The Callistophytales (Pteridospermopsida): reproductively sophisticated Paleozoic gymnosperms. Rev Palaeobot Palynol. 32:103–121.

Rothwell GW. 1982. New interpretations of the earliest conifers. Rev Palaeobot Palynol. 37:7–28.

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, Wojciechowski MF, Hu JM, Khan TS, Brady SG. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol Biol Evol. 17:782–797.

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

Schneider H, Schuettpelz E, Pryer KM, Cranfill R, Magallon S, Lupia R. 2004. Ferns diversified in the shadow of angiosperms. Nature. 428:553–557.

Soltis PS, Soltis DE, Savolainen V, Crane PR, Barraclough TG. 2002. Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. Proc Natl Acad Sci USA. 99:4430–4435.

Stein DB, Conant DS, Ahearn ME, Jordan ET, Kirch SA, Hasebe M, Iwatsuki K, Tan MK, Thomson JA. 1992. Structural rearrangements of the chloroplast genome provide an important phylogenetic link in ferns. Proc Natl Acad Sci USA. 89:1856–1860.

Stevenson DW. 1990. Morphology and systematics of the Cycadales. Mem N Y Bot Gard. 57:8–55.

Stewart WN, Rothwell GW. 1993. Paleobotany and the evolution of plants, 2nd ed. Cambridge: Cambridge University Press. p. 521.

Stewart CN Jr, Via LE. 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques. 14:748–750.

Strauss SH, Palmer JD, Howe GT, Doerksen H. 1988. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. Proc Natl Acad Sci USA. 85:3898–3902.

Sugita M, Murayama Y, Sugiura M. 1994. Structure and differential expression of two distinct genes encoding the chloroplast elongation factor Tu in tobacco. Curr Genet. 25:164–168.

Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Swofford DL, Olsen GJ, Waddell PJ, Hillis DM. 1996. Phylogenetic inference. In: Hillis DM, Moritz C, Mable BK, editors. Phylogenetic inference. Sunderland (MA): Sinauer Associates. p. 407–514.

Tillich M, Lehwark P, Morton BR, Maier UG. 2006. The evolution of chloroplast RNA editing. Mol Biol Evol. 23:1912–1921.

Tsudzuki J, Nakashima K, Tsudzuki T, Hiratsuka J, Shibata M, Wakasugi T, Sugiura M. 1992. Chloroplast DNA of black pine retains a residual inverted repeat lacking rRNA genes: nucleotide sequences of *trnQ*, *trnK*, *psbA*, *trnI* and *trnH* and the absence of *rps16*. Mol Gen Genet. 232:206–214.

Turmel M, Otis C, Lemieux C. 2002a. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci USA. 99:11275–11280.

Turmel M, Otis C, Lemieux C. 2002b. The complete mitochondrial DNA sequence of *Mesostigma viride* identifies this green alga as the earliest green plant divergence and predicts a highly compact mitochondrial genome in the ancestor of all green plants. Mol Biol Evol. 19:24–38.

Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds new light into the closest green algal relatives of land plants. Mol Biol Evol. 23:1324–1338.

Wakasugi T, Hirose M, Horihata T, Tsudzuki T, Kössel H, Sugiura M. 1996. Creation of a novel protein-coding region at the RNA level in black pine chloroplasts: the pattern of RNA editing in the gymnosperm chloroplast is different from that in angiosperms. Proc Natl Acad Sci USA. 93: 8766–8770.

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 USA. 91:9794–9798.

Wang ZO. 2004. A new Permian gnetalean cone as fossil evidence for supporting current molecular phylogeny. Ann Bot. 94:281–288.

Whitfeld PR, Bottemley W. 1983. Organization and structure of chloroplast genes. Annu Rev Plant Physiol. 34:279–310.

Won H, Renner SS. 2006. Dating dispersal and radiation in the gymnosperm *Gnetum* (Gnetales)–clock calibration when outgroup relationships are uncertain. Syst Biol. 55:610–622.

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

William Martin, Associate Editor

Accepted March 15, 2007