The Chloroplast Genome of *Phalaenopsis aphrodite* (Orchidaceae): Comparative Analysis of Evolutionary Rate with that of Grasses and Its Phylogenetic Implications

Ching-Chun Chang,[1] *Hsien-Chia Lin,*[1] *I-Pin Lin,† Teh-Yuan Chow,‡*[2]
Hong-Hwa Chen, Wen-Huei Chen,§ Chia-Hsiung Cheng,‡ Chung-Yen Lin,‖*
Shu-Mei Liu,‡ Chien-Chang Chang,¶ and Shu-Miaw Chaw¶

*Institute of Biotechnology, National Cheng Kung University, Tainan, Taiwan; †Department of Superintendent, Tainan Municipal Hospital, Tainan, Taiwan; ‡Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan; §Department of Life Sciences, National University of Kaohsiung, Kaohsiung, Taiwan; ‖Institute of Information Science, Academia Sinica, Taipei, Taiwan; and ¶Research Center for Biodiversity, Academia Sinica, Taipei, Taiwan

Whether the *Amborella*/*Amborella*-Nymphaeales or the grass lineage diverged first within the angiosperms has recently been debated. Central to this issue has been focused on the artifacts that might result from sampling only grasses within the monocots. We therefore sequenced the entire chloroplast genome (cpDNA) of *Phalaenopsis aphrodite*, Taiwan moth orchid. The cpDNA is a circular molecule of 148,964 bp with a comparatively short single-copy region (11,543 bp) due to the unusual loss and truncation/scattered deletion of certain *ndh* subunits. An open reading frame, *orf91*, located in the complementary strand of the *rrn23* was reported for the first time. A comparison of nucleotide substitutions between *P. aphrodite* and the grasses indicates that only the plastid expression genes have a strong positive correlation between nonsynonymous (K_a) and synonymous (K_s) substitutions per site, providing evidence for a generation time effect, mainly across these genes. Among the intron-containing protein-coding genes of the sampled monocots, the K_s of the genes are significantly correlated to transitional substitutions of their introns. We compiled a concatenated 61 protein-coding gene alignment for the available 20 cpDNAs of vascular plants and analyzed the data set using Bayesian inference, maximum parsimony, and neighbor-joining (NJ) methods. The analyses yielded robust support for the *Amborella*/*Amborella*-Nymphaeales-basal hypothesis and for the orchid and grasses together being a monophyletic group nested within the remaining angiosperms. However, the NJ analysis using K_a, the first two codon positions, or amino acid sequences, respectively, supports the monocots-basal hypothesis. We demonstrated that these conflicting angiosperm phylogenies are most probably linked to the transitional sites at all codon positions, especially at the third one where the strong base-composition bias and saturation effect take place.

Introduction

Recent molecular phylogenetic studies, using one to nine genes from three genomes, have proposed that *Amborella* or the *Amborella*-Nymphaeales clade was the most basal clade among the angiosperms and that the monocots diverged from "some" basal dicots, implying that dicots are paraphyletic (Chaw et al. 1997 [*Amborella* was not sampled in this study]; Mathews and Donoghue 1999; Qiu et al. 1999; see review in Savolainen and Chase 2003; Soltis et al. 2004; D. E. Soltis and P. S. Soltis 2004*a*; P.S. Soltis and D.E. Soltis 2004*b*; Qiu et al. 2005). In stark contrast to this hypothesis, three recent analyses based on 61 common genes of 11, 12, and 14 entire chloroplast genomes (cpDNA), respectively (Goremykin et al. 2003*a*, 2003*b*, 2004), have revealed consistent support for *Amborella* and *Nymphaea* being divergent members of a monophyletic dicot assemblage. Therefore, Goremykin et al. (2004) claimed that the root of angiosperm phylogeny was at the basal monocot-dicot split and that the grasses (Poaceae) were the deepest branch in the phylogeny of the angiosperms.

The discrepancies in the above two hypotheses were focused on three issues: (1) the suitability of sampling only three species of grasses (*Oryza sativa*, *Triticum aestivum*, and *Zea mays*) to represent the monocots in the studies of Goremykin et al., (2) concerns over saturation at the third codon position, and (3) a long-branch attraction (LBA) between the grasses and the outgroup (D. E. Soltis and P. S. Soltis 2004*a*; Soltis et al. 2004; Stefanović, Rice, and Palmer 2004; but see Martin et al. 2005; Lockhart and Penny 2005).

Goremykin et al. (2005) added another species of Poaceae, *Saccharum officinarum*, and a nongrass monocot, *Acorus calamus* (Acoraceae), to their previous cpDNA data set. Their analyses contained support for both the grasses-basal and *Amborella*-Nymphaeales-basal hypotheses. The authors also demonstrated that "the tree recovered under maximum likelihood (ML) is extremely sensitive to model specification, and the best symmetric model is misspecified." Consequently, Goremykin and his colleagues remain agnostic in regard to the phylogeny of basal-most angiosperms. Meanwhile, Leebens-Mack et al. (2005) analyzed an expanded cpDNA data set, notably adding three divergent nongrass monocots, and concluded that the *Amborella*-Nymphaeales clade was basal-most in the phylogeny of the angiosperms, and the positions of the two taxa were dependent on the methods of analysis.

In view of the above controversies concerning the early divergence of angiosperms and the adequacy of sampling, we sequenced the entire cpDNA of *Phalaenopsis aphrodite* Reichb. f., Taiwan moth orchid. This cpDNA data will be a model for the monocot family Orchidaceae, which consists of approximately 30,000 species and is one

[1] These two authors contributed equally to this study.

[2] Present address: Institute of Biotechnology, Central Taiwan University of Science and Technology, Taichung, 406, Taiwan.

Key words: *Phalaenopsis aphrodite*, chloroplast genome, angiosperms, phylogeny, Orchidaceae, Poaceae, substitution rate, molecular evolution.

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

Mol. Biol. Evol. 23(2):279–291. 2006
doi:10.1093/molbev/msj029
Advance Access publication October 5, 2005

Table 1
NCBI Accession Numbers for the Chloroplast Genomes Used in This Study

Classification	Taxon	Accession Number	Reference
Fern			
Psilotaceae	*Psilotum nudum*	NC_003386	Wakasugi et al., unpublished data
Gymnosperms			
Pinaceae	*Pinus koraiensis*	NC_004677	Noh et al., unpublished data
	P. thunbergii	D17510	Wakasugi et al., unpublished data
Angiosperms			
Monocots			
Orchidaceae	*Phalaenopsis aphrodite*	AY916449	Current study
Poaceae			
	Oryza sativa	NC_001320	Hiratsuka et al. (1989)
	O. nivara	NC_005973	Shahid Masood et al. (2004)
	Triticum aestivum	NC_002762	Ogihara et al. (2002)
	Zea mays	NC_001666	Maier et al. (1995)
	Saccharum officinarum	NC_006084	Asano et al. (2004)
Dicots			
Basal dicots			
Amborellaceae	*Amborella trichopoda*	NC_005086	Goremykin et al. (2003*b*)
Nymphaeaceae	*Nymphaea alba*	NC_006050	Goremykin et al. (2004)
Magnoliids			
Calycanthaceae	*Calycanthus floridus* var. *glaucus*	NC_004993	Goremykin et al. (2003*a*)
Core eudicots			
Amaranthaceae	*Spinacia oleracea*	NC_002202	Schmitz-Linneweber et al. (2001)
Araliaceae	*Panax ginseng*	NC_006290	Kim et al. (2004)
Solanaceae	*Atropa belladonna*	NC_004561	Schmitz-Linneweber (2002)
	Nicotiana tabacum	NC_001879	Shinozaki et al. (1986)
Onagraceae	*Oenothera elata*	NC_002693	Hupfer et al. (2000)
Brassicaceae	*Arabidopsis thaliana*	NC_000932	Sato et al. (1999)
Fabaceae	*Lotus japonicus*	NC_002694	Kato et al. (2000)
	Medicago truncatula	NC_003119	Lin et al., unpublished data

of the four largest families of flowering plants (Atwood 1986). Further, we analyzed the substitution rates and patterns among the common genes and introns of *P. aphrodite* and the grasses and correlated their K_s and K_a differences across 65 protein-coding genes that belong to four functional gene groups in an attempt to elucidate the effects resulting from generation time and selection on the plastid genes of different functional groups.

We also tested the two above conflicting hypotheses by comprehensively analyzing a data set with the corresponding 61 protein-coding genes from 20 cpDNAs of vascular plants, including a recently available grass, *Oryza nivara* (wild rice; Shahid Masood et al. 2004), and *P. aphrodite*. Three methods of phylogenetic analyses—Bayesian inference (BI), which has close connection to the ML method but with computational efficiency (Rannala and Yang 1996), maximum parsimony (MP), and neighbor-joining (NJ)—were used to reconstruct the phylogeny of the angiosperms. The robustness of the tree nodes was compared to provide independent evaluations for rooting the angiosperms.

Materials and Methods

Chloroplast DNA Extraction and Genome Sequencing

Leaves of *P. aphrodite* were obtained from seedlings at the four-leaf stage. A voucher specimen has been deposited at the Herbarium of Academia Sinica. Intact chloroplasts were fractionated with step percoll (40%–80%) gradient (Robinson and Downton 1984). DNA was isolated according to acetyl trimethyl ammonium bromide based protocol (Stewart and Via 1993) and sheared into random fragments of 2- to 3-kb pieces using a Hydroshear device (Genomic Solutions Inc., Ann Arbor, Mich.), then cloned into the pBluescriptSK vector to generate a shotgun library. Shotgun clones were propagated, and their plasmids were used as templates for sequencing.

The sequencing reaction was performed using the BigDye terminator cycle sequencing kit (Applied Biosystems, Foster City, Calif.) 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 using the Phred-Phrap programs (Phil Green, University of Washington, Seattle, Wash.). We accumulated 1,728,000 nt, which is about 11.6× coverage. Database searches were conducted with the Blast algorithm in the National Center for Biotechnology Information (NCBI). The tRNAscan program was used to assign the tRNA genes.

Sequence Alignment and Substitution Rate Analysis

The cpDNAs of *P. aphrodite* and the five members of the grass family—*O. sativa*, *O. nivara*, *T. aestivum*, *Z. mays*, and *S. officinarum*—have 65 protein-coding genes in common. Before comparing the substitution rate among them, we made an aligned sequence data set consisting of 10 taxa, which included the cpDNAs of *Calycanthu floridus* var. *glaucus*, *Nicotiana tabacum*, and two species of *Pinus*, *P. koraiensis* and *P. thunbergii* (table 1). *Calycanthus* and *Nicotiana* have relatively low nonsynonymous substitution rates among known cpDNAs, so they were selected as the outgroup for rooting the monocots. The Pamilo-Bianchi-Li

(PBL) method implemented in the software program MEGA 2 (Kumar et al. 2001) was used to calculate the respective number of synonymous (K_s) and nonsynonymous (K_a) substitutions per site in two protein-coding sequences. Divergence of the four rRNA genes and 15 common introns was computed based on the number of substitutions per transitional site (T_s) or per transversional site (T_v) using Kimura's two-parameter (K2P) method (Kimura 1980).

Divergence values between two groups are presented as an average distance ± standard error, obtained from the option "Compute Between Groups Means" in MEGA 2 program. To compare the evolutionary rates of the sampled lineages, Tajima's relative rate test, implemented in MEGA 2 was applied (Tajima 1993). Because the method does not distinguish between K_a and K_s, the first two and the third codon positions were compared separately. Analysis of variance (ANOVA) in the SAS system (SAS Institute Inc., Cary, N.C.) was used to investigate variation in K_s and K_a among different groups of protein-coding genes between *P. aphrodite* and the five grasses.

Reverse Transcriptase–Polymerase Chain Reaction Assay and RNA Filter Hybridization

To verify the expression of *orf91*, total RNAs were extracted from the young leaves of *P. aphrodite* and tobacco using the reagent Trizol (Invitrogen, Carlsbad, Calif.). For reverse transcriptase–polymerase chain reaction (RT-PCR) assay, total RNA was treated with deoxyribonuclease I and then extracted with phenol/chloroform to eliminate any DNA contamination. The resulting RNA was reversely transcribed to make cDNA with a gene-specific primer orf91r (5′-ATGTCTATTTCACCGAGCCT-3′) and the Superscript II reverse transcriptase (Invitrogen, Indianapolis, Ind.). The primer pair, orf91f (5′-TTACCAAAAACACAGGTCTCC-3′) and orf91r, amplified a 249-bp fragment under the following reaction conditions: 94°C for 2 min, followed by 35 cycles of 94°C for 20 s, 55°C for 20 s, and 72°C for 45 s, and ending with a 4-min extension at 72°C.

For the RNA blot, 5 μg of total RNA was electrophoresised in 1% formaldehyde-agarose gel, transferred to nylon membranes, and hybridized at 65°C. To make the RNA probes, the orf91 DNA fragments were PCR amplified and cloned into the pGEM vector to obtain pGEM91. The sense and antisense *orf91* RNA probes were labeled with digoxigenin-11-uridine-5′-triphosphate by in vitro transcription using linear pGEM91 DNA as templates. Blots were washed twice for 15 min at 65°C with 40-mM $NaPO_4$. Hybridized fragments were assayed following the manufacturer's protocol (Roche Applied Science, Indianapolis, Ind.).

Phylogenetic Analysis

Sixty-one protein-coding genes common to the eight complete cpDNAs, including one pine (*P. koraiensis*), three basal dicots (*C. floridus*, *Amborella trichopoda, Nymphaea alba*), two core eudicots (*Atropa belladonna, Panax ginseng*), and two grasses (*S. officinarum*, *O. nivara*), were extracted from the GenBank database and aligned with those of *P. aphrodite* and then integrated with the 12 cpDNA data set of Chaw et al. (2004) to generate two data sets, 20 cpDNAs (table 1) and 15 cpDNAs (i.e., excluding the five grasses).

Because the ML method can generate consistent angiosperm phylogenies from two recent cpDNA data sets (Goremykin et al. 2005; Leebens-Mack et al. 2005) but is far too sensitive to model specification (Goremykin et al. 2005), we applied the BI method instead to analyze our data. 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 is so computationally efficient that it has recently been introduced as a powerful method for reconstructing phylogeny (Holder and Lewis 2003).

Initially, a DNA substitution model for our data sets was selected using Modeltest, Version 3.7 (Posada and Crandall 1998) and the Akaike information criterion. Among the 56 models tested, the general time reversible (GTR) including rate variation among sites (+ G) and invariable sites (+ I) (=GTR + G + I) model was chosen as the best fit to our data sets, followed by the Transversional model + G + I and GTR + G models. For the BI method, we used MrBayes, Version 3.1 (Ronquist and Huelsenbeck 2003). 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 60 (burn-in).

Data sets were also analyzed with MP using the PAUP Version 4 (Swofford 2003) and NJ methods. Distance matrix generated from the GTR + G + I model was also utilized to reconstruct NJ trees in addition to other DNA and amino acid substitution models implemented in MEGA 2. We sought the MP tree with a heuristic search and simple sequence addition, a tree bisection and reconnection branch-swapping algorithm, and a "MulTrees" option in effect. Node confidence was determined by sampling 1,000 bootstrap replicates in both MP and NJ analyses.

Results

Overall Chloroplast Genome Properties and the Finding of *Orf91*

The complete cpDNA molecule of *P. aphrodites* is circular with 148,964 bp (NCBI accession number AY916449) and 64.4% A + T content. It shares a typical structure with the vast majority of land plant cpDNAs—a pair of inverted repeats (IRs) (25,732 bp each) separated by the large single copy (LSC) (85,957 bp) and small single copy (SSC) (11,543 bp) regions. The cpDNA encodes 110 different known genes (Supplementary table 1, Supplementary Material online), including 76 protein-coding genes, 4 rRNA genes, and 30 tRNA genes. In addition, 24 open reading frames (ORFs) were identified with a threshold of 225 bp. RNA editing occurs in the orchid because a C to U conversion was verified in the initiation codon of *rpl2* transcripts.

An ORF, *orf91*, located in the complementary strand of the *rrn23* gene (Supplementary fig. 1*A*, Supplementary Material online) was reported here for the first time. Its amino acid sequence is 81% similar to a hypothetical protein (NCBI accession number ZP_00203428) encoded in the genome of *Anabaena variabilis*. We also observed

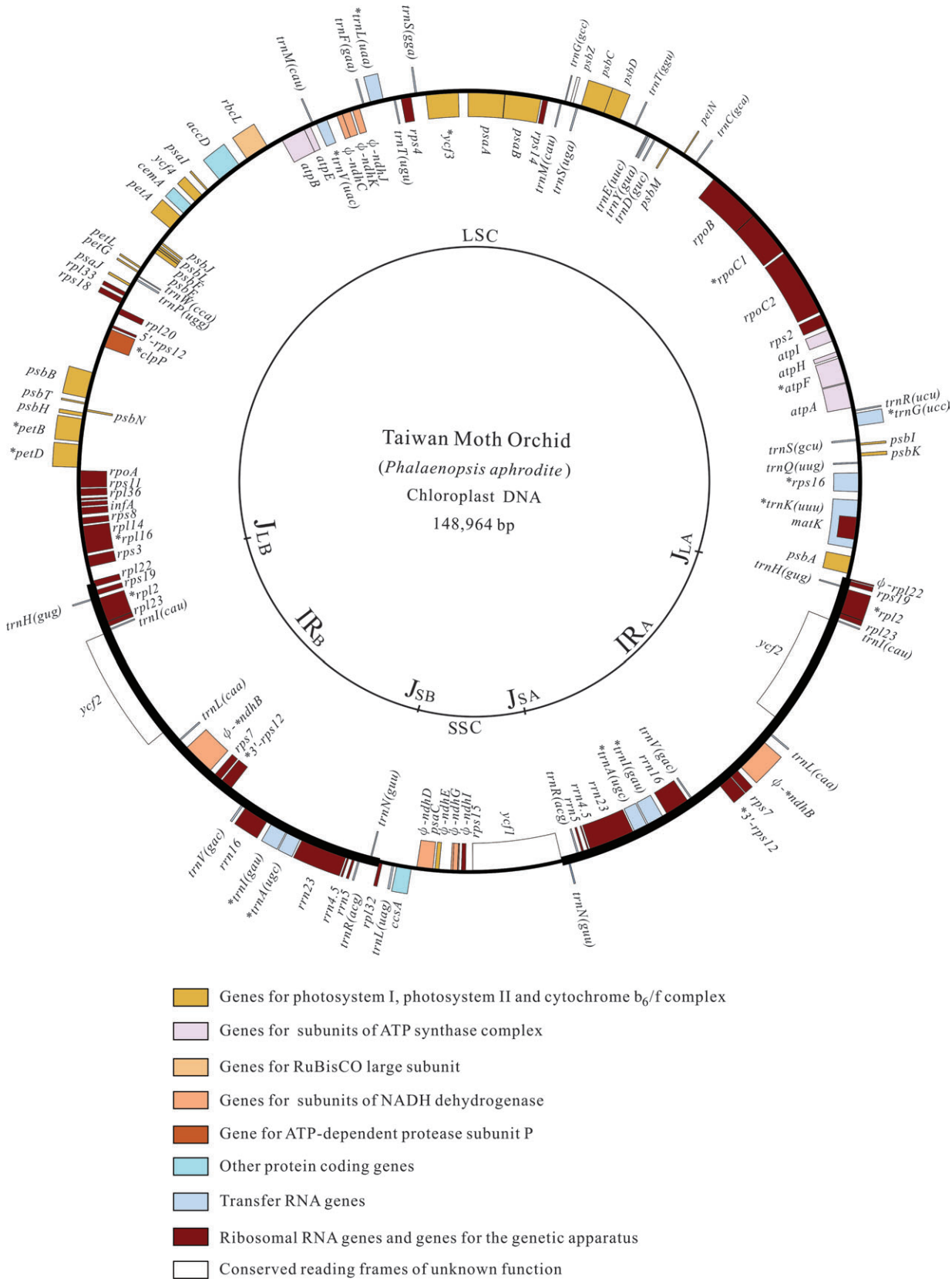

FIG. 1.—Gene map of *Phalaenopsis aphrodite* chloroplast genome. Genes shown inside the circle are transcribed clockwise and those outside are transcribed counterclockwise. Genes belonging to different functional groups are color coded. Asterisks indicate genes containing introns. The four junctions between SC and IR are indicated.

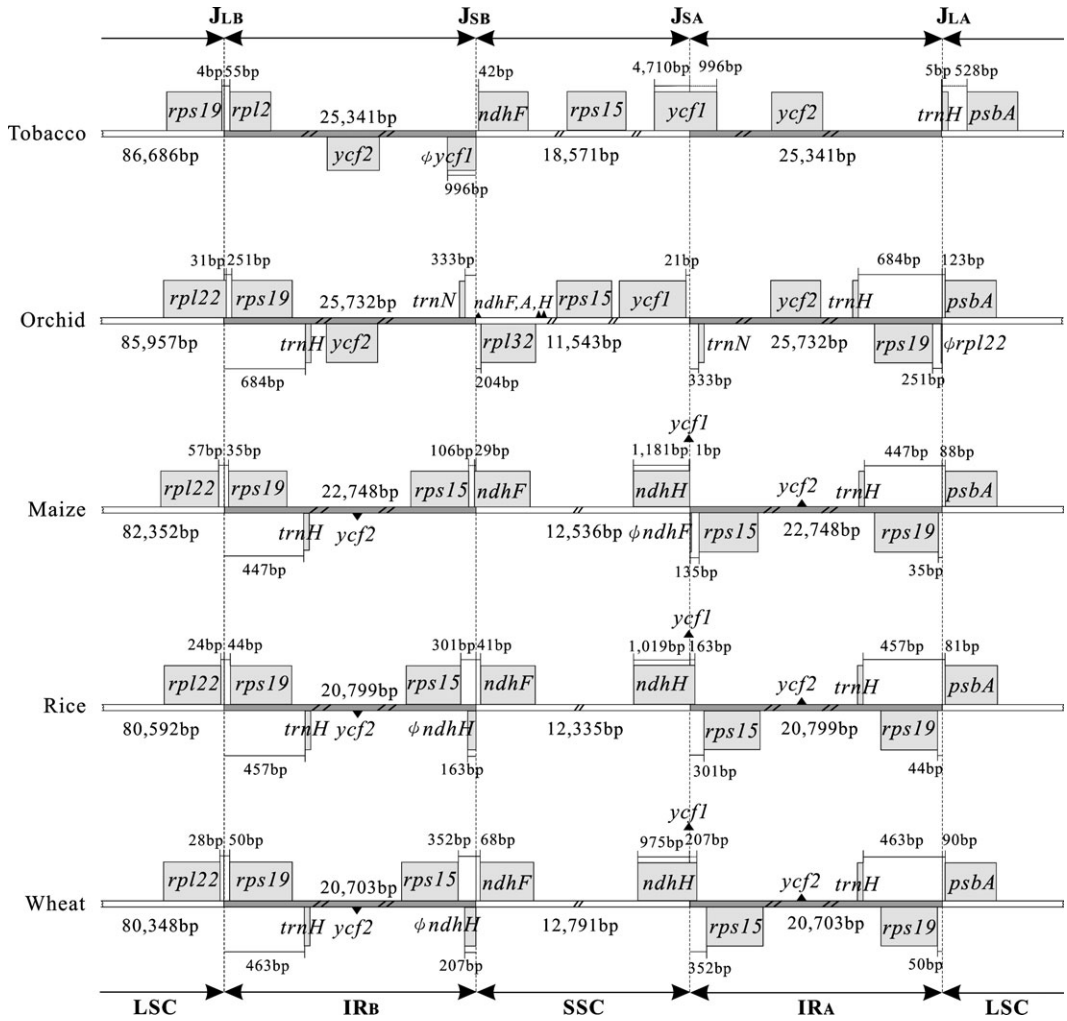

FIG. 2.—Comparison of the junction positions (JLB, JLA, JSB, JSA) between single copies and IR sequences in tobacco, Taiwan moth orchid, and five monocots. Various lengths of *ycf1* or *ndhH* pseudogenes are observed in tobacco, rice, and wheat at the JSB. A fragment of *ndhF* pseudogene (with 29 bp) is found in maize at the JSA. In the orchid 31 bp of *rpl22* pseudogene is located at the JLB. The *rps19-trnH* cluster is located within IR regions in monocots, while in tobacco it is situated in the LSC (data not shown). Note the SSC in monocots being much shorter. Black triangles indicate gene loss.

the *orf91* transcripts in the total RNAs of *P. aphrodite* and tobacco (Supplementary fig. 1*B* and *C*, Supplementary Material online), which testified to the fact that *orf91* was actually transcribed in the chloroplasts. However, further studies are required to determine if the *orf91* transcripts are translated and to characterize the function of the translated protein.

The 11 *ndh* Genes Are Either Lost or Nonfunctional in Taiwan Moth Orchid

Unlike the sequenced cpDNAs of photosynthetic angiosperms that have 11 subunits of *ndh*, in *P. aphrodite* the *ndhA*, *ndhF*, and *ndhH* genes were completely absent. Only the remnants of the *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhG*, *ndhI*, *ndhJ*, and *ndhK* sequences were found. They are either truncated and dispersedly deleted or frameshifted, suggesting that they are nonfunctional. Furthermore, the sequence of *ndhD* begins with ACG rather than ATG, but in our RT-PCR assays, a C to U conversion was not apparent in the *ndhD* transcripts, which indicates that *ndhD* is a pseudogene.

Monocots and Dicots Differ at the Locations of IR and LSC Junctions

The locations of IR and SC junctions have been known to vary among various cpDNAs (Maier et al. 1995; Goulding et al. 1996; Kim and Lee 2004). We compared the exact locations of junctions (viz., JLB, JLA, JSB, and JSA, termed by Shinozaki et al. 1986; fig. 2) and their adjacent genes among tobacco, Taiwan moth orchid, maize, rice, and wheat, in the hope of recognizing a general phylogenetic implication from the available cpDNAs. Figure 2 depicts the JLB of the Taiwan moth orchid to be situated in the coding region of *rpl22*, while in the grasses it is upstream of the gene. In the monocots, JLA is located downstream of the *psbA*. However, the JLB and JLA of tobacco are located upstream of *rps19* and downstream of *trnH*, respectively. In monocots the IR sequences expanded so that two contiguous genes, *trnH* and *rps19*, and their intergenic spacer were encompassed (fig. 2). Although in the IRA of some monocot lineages the 3′ portions of the *rps19* sequences are almost eliminated (such as in *A. calamus*,

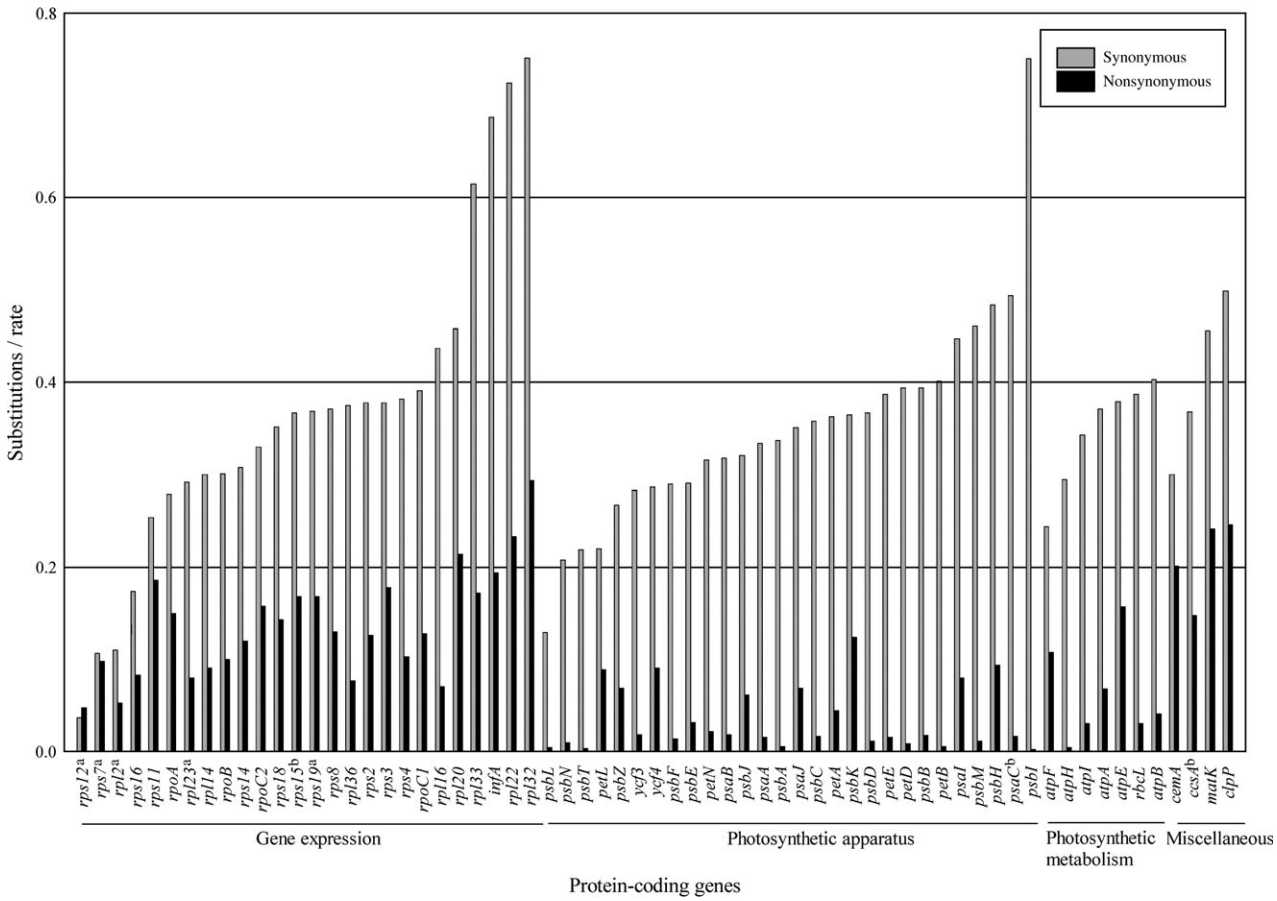

FIG. 3.—Mean K_s and K_a values in each of the 65 common protein-coding genes between Taiwan moth orchid and grasses. Gene order was arranged according to the K_s values within each gene group. Gray and black bars represent K_s and K_a, respectively. [a], IR region and [b], SSC region.

Alisma canaliculatum, and *Pandanus ororayissimus*; Chaw et al., unpublished data), the intergenic spacers preceding the remaining *rps19* sequences are readily alignable with their homologues in the IRB regions. Therefore, we speculate that the *rps19-trnH* cluster was duplicated in the early evolution of monocots (Chaw et al., unpublished data).

In the sampled angiosperms (fig. 2), the downstream sequences of JSB are conserved, with the *ndhF* gene next to it, except in the Taiwan moth orchid in which the gene was lost. The JSA positions in the Taiwan moth orchid and in tobacco resemble each other more than in the grasses in terms of the exact location and upstream gene order in the SSC regions.

SSC Sequences in the Taiwan Moth Orchid and Grasses Are Shorter Than Those in Other Angiosperms

The IR sequences of the grasses are much shorter than those of *P. aphrodite* and tobacco because of the loss of the giant ORF *ycf2* (potentially specifying a protein of 2,290 amino acids in *P. aphrodite*) from each IR region. Furthermore, due to the truncation and scattered deletion of the *ndh* genes in *P. aphrodite* and the loss of the other giant ORF *ycf1* (putatively encoding a protein of 1,816 amino acids in *P. aphrodite*) in the grasses, the SSCs of these two taxa are distinctly shorter than those of *Acorus* and dicots (fig. 2).

K_a Vary Among Groups of Different Functional Genes

Tajima's relative rate test using either the two pines (*P. koraiensis* and *P. thunbergii*), *C. floridus* var. *glaucus*, or tobacco as the outgroup, consistently indicated that substitution rates at all three codon positions of the concatenated 65 gene loci are significantly slower in *P. aphrodite* than in any of the five grasses (all $P < 0.001$).

Figure 3 provides a glimpse into the dynamics of K_s and K_a values across the 65 protein-coding genes between the orchid and grasses. The K_s and K_a are highly variable across genes with the mean difference being 0.36 ± 0.14 and 0.09 ± 0.07 per site, respectively. The K_s and K_a values are the highest in the *rpl32*. In contrast, the *psbI* has the second highest K_s values but the lowest K_a values. Across the 65 protein-coding genes, it is evident that the correlation (Pearson's $r = 0.38$, $P = 0.0016$) between their K_a and K_s values is slight but significant.

To test if the specific gene groups have different influences from generation time effect or selection forces, we divided the 65 genes into four functional groups, gene expression (GE), photosynthetic apparatus (PA), photosynthetic metabolism (PM), and a miscellaneous group (MG). ANOVA indicates significant differences ($P < 0.001$) in mean K_a divergences (md) among the four groups. Duncan's multiple range test also yielded a similar result (for GE vs. MG, md $= -0.072$; GE vs. PA, md $= 0.102$; GE vs. PM,

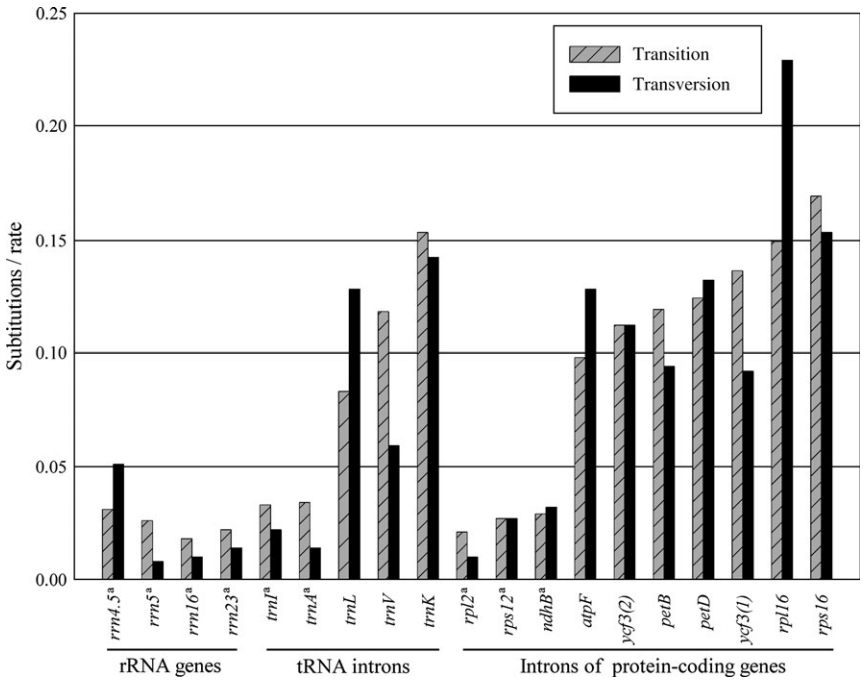

FIG. 4.—Mean nucleotide substitutions in four rRNA genes and 15 introns between cpDNAs of Taiwan moth orchid and grasses. The substitutions are presented as transitions and transversions per site determined by K2P model (Kimura 1980). Number in the parenthesis after an intron name denotes intron order. [a]indicates IR region.

md = 0.074; MG vs. PA, md = 0.174; MG vs. PM, md = 0.146), except for between the PA and PM groups (md = 0.028). However, the MG had the highest mean K_a values, followed by GE, PA, and PM. In contrast, K_s values were not significantly different ($P > 0.05$) among the four groups. Moreover, there is a strong positive correlation (Pearson's $r = 0.76$; $P < 0.0001$) between K_a and K_s in the genes belonging to the GE group but no evidence of significant correlation between the two values in the other three gene groups. In MG, the correlation between the two values is positive but insignificant (Pearson's $r = 0.68$; $P = 0.32$), which is likely due to small sample size. In summary, among the cpDNA protein-coding gene groups the mean K_a differences are more distinct than the mean K_s, and the two values are correlated in the genes that are functionally specific for encoding GE.

K_s of Intron-Containing Genes and T_s of Their Introns Are Significantly Correlated

Figure 4 shows that the T_s and T_v values in the four rRNA genes (4.5S, 5S, 16S, and 23S) of the Taiwan moth orchid and the grasses are relatively low (all less than 0.05 substitutions per site) in comparison with the protein-coding genes in figure 3. As in the rRNA genes, substitution values in the five IR introns of the genes *trnI*, *trnA*, *rpl2*, *rps12*, and *ndhB* were very low. The other 10 shared LSC introns within *trnL*, *trnV*, *trnK*, *atpF*, *ycf3(2)*, *petB*, *petD*, *ycf3(1)*, *rpl16*, and *rps16* had higher T_s and T_v values. We also observed that among the eight common intron-containing protein-coding genes in monocots, the K_s of genes and T_s of their introns are significantly correlated (Pearson's $r = 0.68$, $P = 0.045$), but the K_a of genes and T_a of their introns are not (Pearson's $r = 0.27$; $P = 0.515$).

Amborella and Nymphaeales Are Basal Angiosperms and Monocots Are Monophyletic

After exclusion of unknown sites, difficult-aligned regions, start and stop codons, and all gaps, 38,383 sites were used for comparison and tree reconstruction. The variable sites were 18,022 bp, of which 12,748 were parsimony informative. The BI tree, MP tree, and GTR + G + I-NJ tree (based on the GTR + G + I substitution model) are shown in figure 5*A*1–*C*1, respectively. In these trees, the orchid and grasses always form a clade with 100% bootstrap support, implying that the moncots are monophyletic. The MP method recovered only one tree with 42,090 steps (consistency index [CI], 0.608; retention index [RI], 0.660). The BI and GTR + G + I-NJ trees placed the *Amborella-Nymphaea* clade at the base of the angiosperm phylogeny with robust support, but in the MP tree *Amborella* formed the basal-most clade with 100% bootstrap support, followed by *Nymphaea*, which in turn is sister to the remaining angiosperms with high support (84%). In the GTR + G + I-NJ tree (fig. 5*C*1), the clade containing monocots and core eudicots was moderately supported (77%).

It has been shown that the third codon positions of plastid genes are saturated with substitutions (Goremykin et al. 2003*a*, 2003*b*, 2004; Chaw et al. 2004) and should be excluded from phylogenetic analysis. In this regard, the NJ analysis based on pairwise K_a distance using the PBL model (K_a-NJ tree) was performed. The K_a-NJ tree (fig. 5*D*1) suggests that the dicots and the monocots be separated into two robustly supported monophyletic clades. They diverge at the base of the angiosperms, and *Amborella-Nymphaea* and *Calycanthus* form a clade sister to the eudicots. Moreover, NJ analysis (data not shown) using the combined T_v and T_s distances of the first two

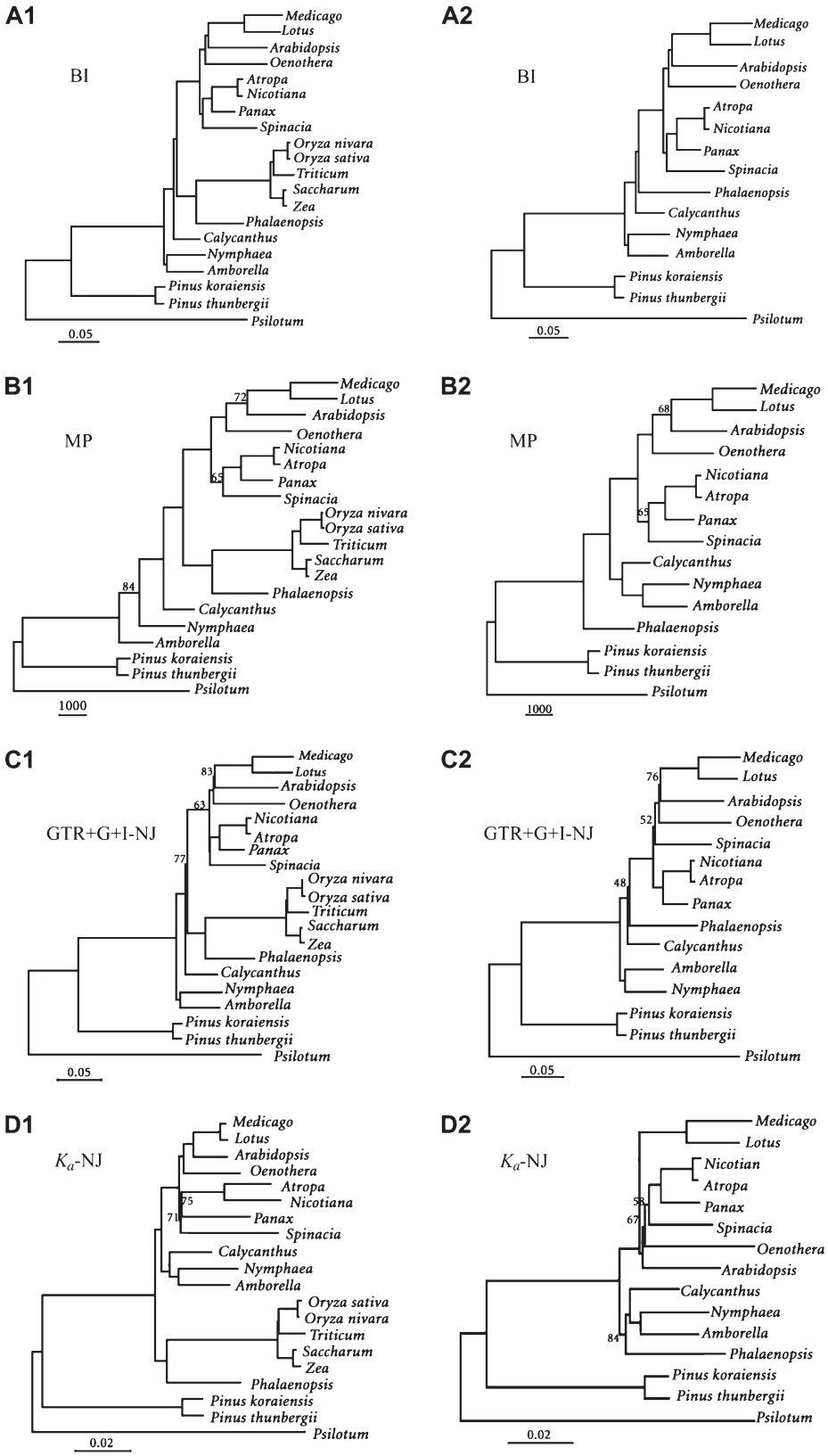

FIG. 5.—(*A*1) BI, (*B*1) MP, (*C*1) GTR + G + I model–based NJ, and (*D*1) K_a-based NJ phylogenetic trees reconstructed from analyses of the nucleotide alignment of 61 protein-coding genes in 20 cpDNA taxa. For *A*2, *B*2, *C*2, and *D*2 phylogenies the five grasses were excluded. *Psilotum* was selected as the outgroup in all trees with the two *Pinus* species used as an internal check. Most nodes on each phylogeny received 100% of the bootstrap replicates, and only values less than 90% were shown for each node. Scale bars denote substitutions per site except for *B*1 and *B*2, where scale bars denote steps.

codon positions—estimated by LogDet, K2P, Tamura's three-parameter, and Tamura-Nei models—or using the amino acid distances estimated by the Dayoff Matrix and Poisson correction models separately, produced topologies identical (data not shown) to that in figure 5*D*1.

Even after removing the grasses from the 20-species data set, angiosperm phylogenies recovered by the BI (fig. 5*A*2) and GTR + G + I-NJ (fig. 5*C*2) methods still resembled figure 5*A*1 and *C*1 in topologies, respectively. Again, the *Amborella-Nymphaea* clade is highly supported and is a sister to the remaining angiosperms. In the MP tree (36,515 steps; CI, 0.648; RI, 0.564) (fig. 5*B*2), the *Phalaenopsis* clade becomes sister to all other angiosperms with 100% bootstrap support. If this MP tree is manually forced to have the same topology as figure 5*B*1, only 75 additional steps will be required (data not shown). K_a-NJ analysis in figure 5*D*2 suggests that *Phalaenopsis* and the (*Calycanthus* (*Amborella*, *Nymphaea*)) form a moderately supported clade, which is sister to the core eudicots. However, the topology of figure 5*D*2 is drastically different from that in figure 5*D*1.

The Core Eudicots Comprise a Monophyletic Clade

All the trees in figure 5 indicate that the sampled eight core eudicots comprise a highly supported monophyletic clade, within which, except for figure 5*D*2, branching orders are congruent with one another (fig. 5*A*1–*D*1, *A*2–*C*2) and there are two sister clades, the Caryophyllales-asterids (*Spinacia* (*Panax* (*Nicotiana*, *Atropa*))) and the eurosids (*Oenothera* (*Arabidopsis* (*Lotus*, *Medicago*))). In each of the clades, the relationships between the taxa are consistent with the conclusions from the analyses of three genes, 18S rDNA, *rbcL*, and *atpB* (Soltis et al. 2000), and four genes, with an additional 26S rDNA (Kim et al. 2004). In figure 5*D*2, *Medicago* and *Lotus* (Fabaceae; table 1) were placed at the base among the sampled core eudicots. This placement is thought to be unreasonable because Fabaceae, a member of the eurosids, becomes an outgroup to the clade containing other two eurosids (*Arabidopsis* and *Oenothera*) and three asteroids (*Panax* (*Nicotiana*, *Atropa*)). We also conducted NJ analysis based on the K_s values (under the PBL model), which produced a branching pattern of the eight core eudicots identical with that of figure 5*D*2. Apparently, the NJ analysis using the K_s distances yielded an unreliable phylogeny within the core eudicots.

Because *Psilotum* has a long branch from the common node of the angiosperms (fig. 5), it was removed from the analyses to avoid potential LBA. Whether or not *Pinus* was designated as the outgroup, the tree topologies and the support at the nodes remained the same in the BI, MP, GTR + G + I-NJ, and K_a-NJ analyses.

Discussion

CpDNA Properties of *P. aphrodites*

The unusual loss and truncation/scattered deletion of *ndh* genes in the SSC region of *P. aphrodites* explain why this region is the smallest among the sequenced photosynthetic angiosperms (fig. 2). The gene order in the LSC region of *P. aphrodites* (fig. 2) is more collinear to those in dicots, such as *Amborella*, *Nymphaea*, *Calycanthus*, and tobacco, than to those in grasses, which include extensive gene rearrangements caused by inversions (Shahid Masood et al. 2004 and cited references). The minisatellite repeat elements contained in the introns of *trnL* of another orchid, *Orchis palustris* (Cafasso 2001), are not found in *P. aphrodites*.

Our preliminary data indicate that in monocots the two IR sequences expand and encompass the *rps19-trnH* cluster, though in certain monocots the *rps19* situated within the IR$_A$ is apparently a truncated pseudogene (Chaw et al., unpublished data). Duplication of this gene cluster most likely represents a major structural difference between the cpDNAs of monocots and dicots.

Our comparative analysis shows that the substitutions per nucleotide site of protein-coding genes between *P. aphrodite* and grasses are highly divergent (fig. 3). In *P. aphrodites* all the protein-coding rRNA and tRNA genes evolve at much slower rates than in the grasses, as indicated in the figure 5*A*1–*D*1. The plastid genes of grasses were reported to have the fastest evolutionary rate (e.g., Clegg at al. 1994) among angiosperms, but our analysis further reveals that the rate differences between the orchid and grasses were primarily a result of the high K_a rate in genes encoding the transcription/translation apparatus and miscellaneous proteins (see Supplementary table 1, Supplementary Material online). The meaning of the high correlation to the evolution of those genes deserves further investigation. *NdhA* and *rpl16* intron sequences were suggested to be good candidates for phylogenetic studies at infrafamilial levels (Kim and Lee 2004). Our analysis indicates that the *trnK* and *rps16* introns may potentially be useful phylogenetically too. Chaw et al. (2005) used the indels of the former to infer the phylogeny of cycad genera.

The cpDNA of *P. aphrodites* contains 24 ORFs (besides *ycf1* and *ycf2*), which encode 77–177 amino acids. Notably, unlike in *P. aphrodites* and another monocot, *A. calamus*, the grass lineage has 20 ORFs and lacks the two giant ORFs, *ycf1* and *ycf 2*, which are present in all known cpDNAs of the dicots and are essential to cell survival in tobacco (Drescher et al. 2000). Therefore, in the grasses these two genes have probably been transferred to the nuclear genome. Further work on the functional genomics of the 24 ORFs is required to elucidate the evolution of such diversified ORFs in *P. aphrodites*.

The Ancestral Plastid *ndh* Copies of *P. aphrodites* May Have Been Transferred to the Nucleus

Chloroplasts were once free-living cyanobacteria and now retain preserved remnants of eubacterial genomes (Martin et al. 1998), but some of their genes have been transferred to the nucleus during evolution (Palmer 1985, 1991; Martin et al. 2002). The *ndh* genes that encode the 11 subunits of the nicotinamide adenine dinucleotide dehydrogenase complex are involved in the cyclic electron flow of photosystem I and chloroplast respiration (Burrows et al. 1998). We did not detect RNA editing in the *ndhD*, *ndhJ*, and *ndhK* transcripts, implying that the eight variously truncated and frameshifted *ndh* genes are probably pseudogenes. Loss of all functional *ndh* genes was reported in

the cpDNAs of the parasitic plant *Epifagus virginiana* (Wolfe, Morden, and Palmer 1992) and in *P. thunbergii* (Wakasugi et al. 1994). Some orchids also do not have the plastid *ndhF* (Neyland and Urbatsch 1996). We have confirmed the in-frame sequences of the lost plastid genes *ndhA*, *ndhF*, and *ndhH* in the total DNA of *P. aphrodites* by PCR assay (unpublished data), implying that the ancestral functional *ndh* copies of the plastid may have been transferred to the nuclear genome. Further studies will unravel the details of the lost *ndh* genes in the plastid of *P. aphrodites*.

Correlation of K_a and K_s Values Is Significantly Positive in the Genes Encoding Expression

Ideally, rate comparison should be made in a multilocus context (Clegg et al. 1994; Gaut et al. 1997). For example, the minimum generation time effect should influence rates in all loci within a genome, and therefore "evidence of generation time effects should be consistent across loci" (Gaut et al. 1997). The evolutionary force affecting multiple genes in plastid nucleotide substitution rates among monocot lineages remains unexplored. The complete cpDNA of the orchid provides an opportunity for just such a comparison. We hereby extend the previous work of Gaut et al. (1997) and compare the K_s and K_a values to examine rate dynamics at the shared 65 protein-coding genes in *Phalaenopsis* and the grass lineages. The comparisons across 2 or 33 plastid genes have indicated that K_s and K_a are uncoupled (uncorrelated) (Gaut et al. 1997; Muse and Gaut 1997). However, our data clearly indicate that in monocots the K_s and K_a values are slightly but significantly correlated (Pearson's $r = 0.38$, $P = 0.0016$) across 65 plastid protein-coding genes, and this correlation is even strongly significant in the genes encoding expression.

The generation time of *S. officinarum* (sugarcane) is 1–2 years (Australian Government 2004), which is longer than the other four selected annual grasses (Clayton and Renvoize 1986). We observed that both K_s and K_a values of sugarcane are not significantly different from those of rice, wheat, and maize (data not shown). In contrast, *P. aphrodite* is a perennial epiphytic orchid and takes 2–3 years to set flowers and seeds (Chang et al. 2000). Because in the orchid and grasses a strong positive correlation exists between K_s and K_a values of their plastid expression genes, it appears that there is some evidence for generation time effects. Furthermore, because the PA and PM have relatively low (or conserved) K_a, we consider that the selection may dominantly take effect on the nonsynonymous sites of these genes.

The *Amborella*-Nymphaeales-Basal Hypothesis Receives More Support

With the BI, MP, and GTR + G + I-NJ analyses in this study, the ML analysis in both Goremykin et al. (2005) and Leebens-Mack et al. (2005), and the MP and Hasegawa, Kishino, Yano (HKY)-NJ analyses in Leebens-Mack et al. (2005), which held that the concatenated cpDNA data sets do not support the monocot/grasses-basal hypothesis, the approaches which still support this hypothesis are the K_a-NJ analyses in Goremykin et al. (2003*a*, 2003*b*, 2004) and this study and the MP and GTR-NJ analyses in Goremykin et al. (2005). Overall then, the growing cpDNA databases provide stronger support for the *Amborella-Nymphaea*/*Amborella*-basal than the grasses-basal hypothesis. If the latter hypothesis does stand the test of time, then considerable reinterpretation of the evolution of dicots and nongrass monocots will be required.

The Third Codon Positions Are Informative But with "Misinformative" Transitional Sites

Because one (or perhaps none) of the above two hypotheses represents the true phylogeny of angiosperms, we would like to echo the comment of Leebens-Mack et al. (2005) that "inconsistencies among results derived from different approaches, models or datasets should be examined and explained rather than ignored." Saturation at the third codon position of plastid genes has previously been documented (e.g., Goremykin et al. 1997, 2003*a*, 2003*b*, 2004; Chaw et al. 2004), and amino acid composition bias in chloroplast proteins can strongly affect the plastid genome phylogeny (Martin et al. 2002). Nevertheless, Leebens-Mack et al. (2005, fig. 2) claimed that saturation did not seriously bias distance estimates because they mapped a linear relationship between the genetic divergences (HKY distance) occurring at the first two and the third codon positions. However, a close inspection of their figure 2 proves that the HKY distances at the third codon position increase much faster than at the first two. That is, if the *x* axis (first two positions) and *y* axis (third position) are at the same scale, a regression line for those distance dots would incline to the *y* axis. Hence, in the data set of Leebens-Mack et al. (2005) substitutions at the third codon position likely have the tendency to be saturated.

We found that the K_s values of intron-containing genes and the T_s values of their introns are moderately correlated. Does our finding connote that evolutionary constraint actually occurs at the third codon position? Is it appropriate to simply exclude the third codon position from the distance-based NJ analysis or only use the K_a distances in NJ analysis for plastid genes? To test this assumption, the K2P-based NJ analysis was used to analyze the two new data sets: (1) the first two and (2) the third codon positions of our 20 cpDNA data set (or with the five grasses removed) separately. The T_v distances of both new data sets supported the *Amborella*-Nymphales-basal hypothesis. In contrast, the T_s distances of the first two positions yielded a tree similar to the K_a-NJ tree of figure 5*D*1, which sustains the monocot basal hypothesis. The T_s distances of the third position data set yielded a tree identical to that of figure 5*D*2, in which, as pointed out before, the lineage relationships within the sampled eight core eudicots were unreasonable. Hence, it is very obvious that transitional and transversional sites at the first two codon positions contain conflicting phylogenetic signals. Furthermore, transitional sites at the third codon position consist of misinformative characters that can yield incorrect phylogenies.

The A-T is very rich in plastid genes and particularly strong at the third codon position (Martin et al. 2002; Goremykin et al. 2003*a*, 2003*b*, 2004; Chaw et al. 2004), so that there are base-composition biases in the concatenated

plastid protein-coding sequence data, and "most methods will incorrectly group OTUs with similar base composition" (Li and Graur 1999). Moreover, when there are many homoplasies in the data, MP methods will yield erroneous trees (Li and Graur 1999). Therefore, cautious selection of DNA substitution models and methods is a prerequisite for analyzing cpDNA data set. Finally, we would like to defer to the comment of Goremykin et al. (2005) that "model misspecification may also arise due to fitting a single model to an alignment that contains many concatenated genes" because a highly variable substitution pattern was observed among the cpDNA genes of the sampled monocots.

Concerning the discrepancies in the resolution of angiosperm phylogeny, such issues as taxon sampling, the fitness of model to the data, and model specification have been comprehensively discussed and reviewed (e.g., Soltis et al. 2004; Goremykin et al. 2005; Leebens-Mack et al. 2005; Martin et al. 2005; and reference herein). However, analysis of a larger data set with quantities of genes and taxa is needed to unambiguously address the deepest questions in angiosperm phylogeny.

Supplementary Material

Supplementary tables 1–3 and figure 1*A–C* are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Supplementary Figure 1. Detection of *orf91* RNA transcripts: (*A*) Relative position of *orf91* in chloroplast genome. Gene size is not to scale. Probe position and size are indicated. (*B*) RT-PCR assay. Lane 1, *Phalaenopsis aphrodite* cDNA; Lane 2, *P. aphrodite* RNA; Lane 3, tobacco cDNA; Lane 4, tobacco RNA. DNA sizes are indicated to the left. The expected 249-bp DNA fragments were only amplified from the cDNA of *Phalaenopsis* and tobacco. (*C*) RNA gel blot analysis: 5 μg of total RNA from *P. aphrodite* (Lanes 1 and 3) and tobacco (Lanes 2 and 4) was loaded. Probes are indicated below each gel. The *rrn23* transcripts of expected 2,808 bp were detected when the *orf91* sense probe was used as a control. In contrast, 2,050-bp RNA transcripts in the total RNA were detected in *Phalaenopsis* and tobacco when an *orf91* antisense probe was used.

Acknowledgments

We thank Yi-Fen Chiu and Chiao-Lei Cheng for their technical assistance, Ai-Ling Hou for her assistance with using the SAS system, and Sean Turner for his help with running the Modeltest program. We are grateful to Kenneth H. Wolfe, Rui-Jiang Wang, and an anonymous reviewer for their critical comments and suggestions on an early version of the manuscript. This work was supported by a National Science Council grant (NSC92-2317-B006-004) to C.C.C. and in part by a grant from the Research Center for Biodiversity, Academia Sinica to S.M.C.

Literature Cited

Asano, T., T. Tsudzuki, S. Takahashi, H. Shimada, and K. Kadowaki. 2004. Complete nucleotide sequence of the sugarcane (*Saccharum officinarum*) chloroplast genome: a comparative analysis of four monocot chloroplast genomes. DNA Res. **11**:93–99.

Atwood, J. T. 1986. The size of the Orchidaceae and the systematic distribution of epiphytic orchids. Selbyana **9**:171–186.

Australian Government. 2004. The biology and ecology of sugarcane (*Saccharum* spp. hybrids) in Australia. (http://www.ogtr.gov.au/rtf/ir/biologysugarcane.rtf).

Burrows, P. A., L. A. Sazanov, Z. Svab, P. Maliga, and P. J. Nixon. 1998. Identification of a functional respiratory complex in chloroplasts through analysis of tobacco mutants containing disrupted plastid *ndh* genes. EMBO J. **17**:868–876.

Cafasso, D., G. Pellegrino, A. Musacchio, A. Widmer, and S. Cozzolino. 2001. Characterization of a minisatellite repeat locus in the chloroplast genome of *Orchis palustris* (Orchidaceae). Curr. Genet. **39**:394–398.

Chang, S.-B., W.-H. Chen, H.-H. Chen, Y.-M. Fu, and Y.-S. Lin. 2000. RFLP and inheritance patterns of chloroplast DNA in intergenic hybrids of *Phaalaenopsis* and *Doritis*. Bot. Bull. Acad. Sin. **41**:219–223.

Chaw, S.-M., C.-C. Chang, H.-L. Chen, and W.-H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J. Mol. Evol. **58**: 1–18.

Chaw, S.-M., T. W. Walters, C.-C. Chang, S.-H. Chen, and S.-H. Hu. 2005. Phylogeny of cycad genera inferred from chloroplast *matK* gene, *trnK* intron, and nuclear *rDNA* ITS region. Mol. Phylogenet. Evol. **37**:214–234.

Chaw, S.-M., A. Zharkikh, H. M. Sung, T. C. Lau, and W. H. Li. 1997. Molecular phylogeny of extant gymnosperms and seed plant evolution: analysis of nuclear 18S rRNA sequences. Mol. Biol. Evol. **14**:56–68.

Clayton, W. D., and S. A. Renvoize. 1986. Genera graminurn: grasses of the world. Kew Bull. **13**:1–389.

Clegg, M. T., B. S. Gaut, G. H. Learn Jr., and B. R. Morton. 1994. Rates and patterns of chloroplast DNA evolution. Proc. Natl. Acad. Sci. USA **91**:6795–6801.

Drescher, A., S. Ruf, T. Calsa, H. Carrer, and R. Bock. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. **22**:97–104.

Gaut, B. S., L. G. Clark, J. F. Wendel, and S. V. Muse. 1997. Comparisons of the molecular evolutionary process at rbcL and ndhF in the grass family (Poaceae). Mol. Biol. Evol. **14**:769–777.

Goremykin, V., S. Hansmann, and W. Martin. 1997. Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. Plant Syst. Evol. **206**:337–351.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2003*a*. The chloroplast genome of the basal angiosperm *Calycanthus fertilis*—structure and phylogenetic analysis. Plant Syst. Evol. **242**:119–135.

———. 2003*b*. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol. Biol. Evol. **20**:1499–1505.

———. 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol. Biol. Evol. **21**:1445–1454.

Goremykin, V. V., B. Holland, K. I. Hirsch-Ernst, and F. H. Hellwig. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol. Biol. Evol. **22**: 1813–1822.

Goulding, S. E., R. G. Olmstead, C. W. Morden, and K. H. Wolfe. 1996. Ebb and flow of the chloroplast inverted repeat. Mol. Gen. Genet. **252**:195–206.

Hiratsuka, J., H. Shimada, R. Whittier et al. (10 co-authors). 1989. The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. **217**:85–94.

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

Hupfer, H., M. Swiatek, S. Hornung, R. G. Herrmann, R.M. Maier, W. L. Chiu, and B. Sears. 2000. Complete nucleotide sequence of the *Oenothera elata* plastid chromosome, representing plastome I of the five distinguishable euoenothera plastomes. Mol. Gen. Genet. **263**:581–585.

Kato, T., T. Kaneko, S. Sato, Y. Nakamura, and S. Tabata. 2000. Complete structure of the chloroplast genome of a legume, *Lotus japonicus*. DNA Res. **7**:323–330.

Kim, K. J., and H. L. Lee. 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.

Kim, S., D. E. Soltis, P. S. Soltis, M. J. Zanis, and Y. Suh. 2004. Phylogenetic relationships among early-diverging eudicots based on four genes: were the eudicots ancestrally woody? Mol. Phylogenet. Evol. **31**:16–30.

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

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

Leebens-Mack, J. H., L. A. Raubeson, L. Cui, J. V. Kuehl, M. H. Fourcade, T. W. Chumley, J. L. Boore, R. K. Jansen, and C. W. dePamphilis. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol. Biol. Evol. **22**:1948–1963.

Li, W.-H., and D. Graur. 1999. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.

Lockhart, P. J., and D. Penny. 2005. The place of *Amborella* within the radiation of angiosperms. Trends Plant Sci. **10**:477–483.

Maier, R. M., K. Neckermann, G. L. Igloi, and H. Kossel. 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.

Martin, W., O. Deusch, N. Stawski, N. Grünheit, and V. Goremykin. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. **10**:203–209.

Martin, W., T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. 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., B. Stoebe, V. Goremykin, S. Hapsmann, M. Hasegawa, and K. V. Kowallik. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature **393**:162–165.

Mathews, S., and M. J. Donoghue. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science **286**:947–950.

Muse, S. V., and B. S. Gaut. 1997. Comparing patterns of nucleotide substitution rates among chloroplast loci using relative ratio test. Genetics **146**:393–399.

Neyland, R., and L. E. Urbatsch. 1996. The *ndhF* chloroplast gene detected in all vascular plant divisions. Planta **200**: 273–277.

Ogihara, Y., K. Isono, T. Kojima et al. (19 co-authors). 2002. Structural features of a wheat plastome as revealed by complete sequencing of chloroplast DNA. Mol. Genet. Genomics **266**:740–746.

Palmer, J. D. 1985. Comparative organization of chloroplast genomes. Annu. Rev. Genet. **19**:325–354.

———. 1991. Plastid chromosome: structure and evolution. Pp. 5–53 *in* I. K. Vasil and L. Bogorad, eds. Cell culture and somatic cell genetics in plants, Vol. 7A. The molecular biology of plastids. Academic Press, San Diego, Calif.

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

Qiu, Y. L., O. Dombrovska, J. H. Lee et al. (19 co-authors). 2005. Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. Int. J. Plant Sci. **166**:815–842.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, Z. Chen, V. Savolainen, and M. W. Chase. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature **402**:404–407.

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

Robinson, S. P., and W. J. Downton. 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., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572–1574.

Sato, S., Y. Nakamura, T. Kaneko, E. Asamizu, and S. Tabata. 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res. **6**:283–290.

Savolainen, V., and M. W. Chase. 2003. A decade of progress in plant molecular phylogenetics. Trends Genet. **19**:717–724.

Schmitz-Linneweber, C., R. M. Maier, J. P. Alcaraz, A. Cottet, R. G. Herrmann, and R. Mache. 2001. The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol. Biol. **45**:307–315.

Schmitz-Linneweber, C., R. Regel, T. G. Du, H. Hupfer, R. G. Herrmann, and R. M. Maier. 2002. The plastid chromosome of *Atropa belladonna* and its comparison with that of *Nicotiana tabacum*: the role of RNA editing in generating divergence in the process of plant speciation. Mol. Biol. Evol. **19**:1602–1612.

Shahid Masood, M., T. Nishikawa, S. Fukuoka, P. K. Njenga, T. Tsudzuki, and K. Kadowaki. 2004. The complete nucleotide sequence of wild rice (*Oryza nivara*) chloroplast genome: first genome wide comparative sequence analysis of wild and cultivated rice. Gene **340**:133–139.

Shinozaki, K., M. Ohme, M. Tanaka, T. Wakasugi, N. Hayashida, and M. Sugiura. 1986. The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression. EMBO J. **5**:2043–2049.

Soltis, D. E., V. A. Albert, V. Savolainen et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. **9**:477–483.

Soltis, D. E., and P. S. Soltis. 2004*a*. *Amborella* not a "basal angiosperm"? Not so fast. Am. J. Bot. **91**:997–1001.

Soltis, P. S., and D. E. Soltis. 2004*b*. The origin and diversification of angiosperms. Am. J. Bot. **91**:1614–1626.

Soltis, D. E., P. S. Soltis, M. W. Chase et al. (16 co-authors). 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. Bot. J. Linn. Soc. **133**:381–461.

Stefanović, S., D. W. Rice, and J. D. Palmer. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? BMC Evol. Biol. **4**:35.

Stewart, C. N. Jr., and L. E. Via. 1993. A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. Biotechniques **14**:748–750.

Swofford, D. L. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods).Version 4. Sinauer Associates, Sunderland, Mass.

Tajima, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **10**:677–688.

Wakasugi, T., J. Tsudzuki, S. Ito, K. Nakashima, T. Tsudzuki, and M. Sugiura. 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.

Wolfe, K. H., C. W. Morden, and J. D. Palmer. 1992. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. Proc. Natl. Acad. Sci. USA **89**: 10648–10652.

William Martin, Associate Editor

Accepted September 23, 2005