American Journal of Botany 87(11): 1712–1730. 2000.

UTILITY OF 17 CHLOROPLAST GENES FOR INFERRING THE PHYLOGENY OF THE BASAL ANGIOSPERMS[1]

SEAN W. GRAHAM[2,3] AND RICHARD G. OLMSTEAD[2]

[2]Department of Botany, Box 355325, University of Washington, Seattle, Washington, 98195-5325 USA

Sequences from 14 slowly evolving chloroplast genes (including three highly conserved introns) were obtained for representative basal angiosperm and seed-plant taxa, using novel primers described here. These data were combined with published sequences from *atpB*, *rbcL*, and newly obtained sequences from *ndhF*. Combined data from these 17 genes permit sturdy, well-resolved inference of major aspects of basal angiosperm relationships, demonstrating that the new primers are valuable tools for sorting out the deepest events in flowering plant phylogeny. Sequences from the inverted repeat (IR) proved to be particularly reliable (low homoplasy, high retention index). Representatives of *Cabomba* and *Illicium* were the first two successive branches of the angiosperms in an initial sampling of 19 exemplar taxa. This result was strongly supported by bootstrap analysis and by two small insertion/deletion events in the slowly evolving introns. Several paleoherb groups (representatives of Piperales) formed a strongly supported clade with taxa representing core woody magnoliids (Laurales, Magnoliales, and Winteraceae). The monophyly of the sampled eudicots and monocots was also well supported. Analyses of three major partitions of the data showed many of the same clades and supported the rooting seen with all the data combined. While *Amborella trichopoda* was supported as the sister group of the remaining angiosperms when we added *Amborella* and *Nymphaea odorata* to the analysis, a strongly conflicting rooting was observed when *Amborella* alone was added.

Key words: *Amborella*; *atpB*; basal angiosperm phylogeny; chloroplast introns; long-branch attraction; NADH dehydrogenase genes; Photosystem II genes; primer design; *rbcL*; ribosomal protein genes.

The pattern of early evolutionary diversification of the flowering plants has been one of the most outstanding unresolved problems in plant biology. Major contributions to our understanding of early angiosperm relationships have come from recent studies of a variety of morphological and anatomical characters in living and fossil taxa (e.g., Donoghue and Doyle, 1989; Loconte and Stevenson, 1991; Taylor and Hickey, 1992; Hickey and Taylor, 1996; Loconte, 1996; Tucker and Douglas, 1996; Endress and Igersheim, 1997; Igersheim and Endress, 1997, 1998; Nandi, Chase and Endress, 1998). Unfortunately, these characters typically provide only weak or ambiguous support for deep angiosperm relationships when used in phylogenetic analysis (e.g., Crepet, 1998; Doyle, 1998). This is a consequence of the relatively low number of morphological characters currently available, but is also because it can be difficult to determine character homology across the major seed-plant groups (J. A. Doyle, personal communication, University of California, Davis; Frohlich, 1999). An additional major problem is that the angiosperms are by far the largest group of living land plants, and so an ''exemplar'' taxon sampling approach is unavoidable. Fortunately, most angiosperms fall in a few species-rich clades (the monocots and eudicots in particular; Chase et al., 1993), so relatively few taxa may suffice to sample adequately the phylogenetic diversity of the most deeply branching lineages.

[1] Manuscript received 22 April 1999; revision accepted 29 February 2000.

The authors thank Patrick Reeves for technical assistance. James Doyle, Patrick Reeves, and Michael Sanderson, and two anonymous reviewers provided critical readings of the manuscript. We are grateful to two reviewers for their strong encouragement to add *Amborella* to the study. Mark Chase, Sara Hoot, and Vincent Savolainen shared *atpB* sequences prior to publication, and Mark Chase, Dan Crawford, Sara Hoot, Thomas Lemieux, and Don Les contributed DNA samples and plant material. This work was funded by NSF grant DEB 9727025.

[3] Author for reprint requests, current address: Department of Biological Sciences, Biological Sciences Centre, University of Alberta, Edmonton, Alberta, Canada T6G 2E9.

The fossil record of the flowering plants extends back at least 130 million years (see Crane, 1993; Doyle and Donoghue, 1993; Crane, Friis, and Pedersen, 1995). Several lines of molecular evidence suggest an even earlier origin of the crown angiosperm group (reviewed in Sytsma and Baum, 1996; Li, 1997), and the available paleobotanical data do not rule out an extended period of unrecorded angiosperm evolution (Crane, 1993). Regardless of this uncertainty, very long branches subtend the earliest diverging or ''most basal'' lineages of the angiosperms in phylogenetic analysis. In contrast, a preponderance of relatively short internodes around the base of the angiosperms (e.g., Chase, 1993) suggests a relatively rapid diversification of many of the extant basal lineages. The combination of deep lineages with short internal branches is known to have the potential for yielding strongly misleading results, the phenomenon of statistical inconsistency, or ''long-branch attraction'' (Felsenstein, 1978; Hendy and Penny, 1989).

Long branches may be difficult or impossible to divide by additional taxon sampling for at least some basal lineages, such as Amborellaceae and Ceratophyllaceae, that are represented by only one or a few extant taxa. Swofford and Poe (1999) caution that adding taxa to break up long branches will not always improve the consistency of phylogenetic estimation, the ability to converge on the correct answer with increasing amounts of data. Compounding the problem of long branches in basal angiosperm lineages, the extant seed-plant groups, the angiosperms, conifers, cycads, *Ginkgo* and Gnetales, represent phylogenetically disjunct remnants of a more ancient radiation. Correspondingly long branches separate the major living seed-plant groups. One solution to circumventing long-branch attraction, in the absence of living lineages that could span this divide, may be to employ conservatively evolving characters (Felsenstein, 1983).

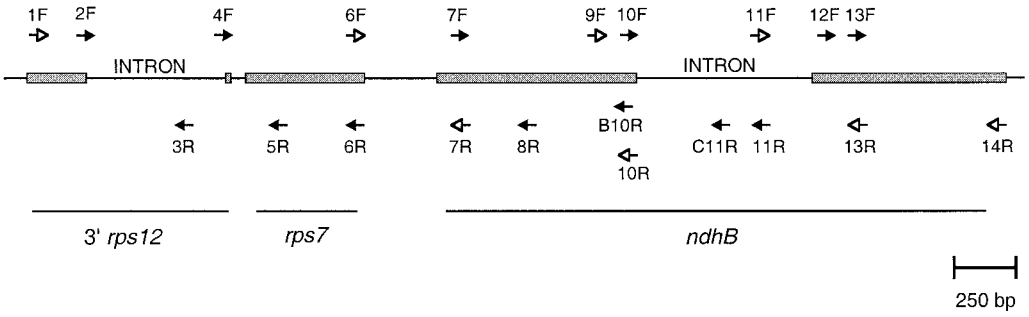
Sampling error on short internal branches is an additional, poorly recognized source of ambiguous and misleading phy-

TABLE 1. Sources of angiosperm and other seed plant DNA sequences employed in the current study.

Exemplar species	Voucher for sequences obtained for the current study	Previously published sequences used here, with GenBank accession no.[a]
Amborellaceae		
Amborella trichopoda	*Mark P. Simmons* 1846 (GH)	*rbcL* (GBAN-L12628)
Nymphaeales		
Cabomba caroliniana	*Les s.n.* (CONN)	*rbcL* (GBAN-M77027)
Nymphaea odorata		*rbcL* (GBAN-M77034); *atpB* (GBAN-AJ235544); 15 genes (GBAN-AF188851-AF188856)
Illiciaceae		
Illicium parviflorum	*Naczi 2784* (MICH)	*rbcL* (GBAN-L12652); *atpB* (GBAN-U86385)
Laurales		
Calycanthus floridus	Matthaei Bot. Gard., Ann Arbor, MI, no voucher; (*RGO 307* DNA)	*rbcL* (GBAN-L14291); *atpB* (GBAN-AJ235422)
Magnoliales		
Liriodendron tulipifera	*RGO 97-9* (WTU)	*rbcL* (GBAN-X54346); *atpB* (GBAN-AJ235522)
Piperales		
Asarum canadense	Ann Arbor, MI, no voucher; (*RGO 748* DNA)	*rbcL* (GBAN-L14290); *atpB* (GBAN-U86383)
Lactoris fernandeziana	*Stuessy et al. 11591* (OS)	*rbcL* (GBAN-L08763); *atpB* (GBAN-AJ235515)
Saururus cernuus	*RGO 88-006* (WTU)	*rbcL* (GBAN-L14294)
Winteraceae		
Drimys winteri	*RGO 97-13* (WTU)	*rbcL* (GBAN-L01905); *atpB* (GBAN-AF093425)
Ceratophyllaceae		
Ceratophyllum demersum	*Les s.n.* (CONN)	*rbcL* (GBAN-D89473); *atpB* (GBAN-AJ235430)
Monocots		
Acorus calamus	Denver Botanic Gard., CO, no voucher; (*RGO 97-149* DNA)	*rbcL* (GBAN-D28865); *atpB* (GBAN-AJ235381)
Oryza sativa		Complete chloroplast genome (GBAN-X15901)
Dioscorea bulbifera	EPO Biology, U. Colorado, Boulder, CO, no voucher; (*RGO 97-151* DNA)	*rbcL* (GBAN-D28327)
Zea mays		Complete chloroplast genome (GBAN-X86563)
Eudicots		
Cercidiphyllum japonicum	*RGO 90-016* (WTU)	*rbcL* (GBAN-L11673); *atpB* (GBAN-AF092112)
Nicotiana tabacum		Complete chloroplast genome (GBAN-Z00044; GBAN-S54304)
Trochodendron araliodes	*RGO 97-8* (WTU)	*rbcL* (GBAN-L01958); *atpB* (GBAN-AF093423)
Seed plant outgroups		
Ginkgo biloba	*SWG 97-IV (1)* (WTU)	*rbcL* (GBAN-D10733); *atpB* (GBAN-AJ235480)
Gnetum gnemon	*RGO 97-16* (WTU)	*rbcL* (GBAN-U72819)
Pinus thunbergii		Complete chloroplast genome (GBAN-D17510)

[a] The prefix GBAN–has been added to link the online version of *American Journal of Botany* to GenBank, but is not part of the actual accession number.

logenetic inference (Rodrigo et al., 1993; Page, 1996; Graham et al., 1998). The accuracy of phylogenetic inference may be more easily improved by adding characters to detect changes on short internal branches, than by adding additional taxa (Swofford and Poe, 1999). However, if slowly evolving regions are to be used to address basal angiosperm relationships, more characters in total must be sampled to assure sufficient resolution of deep, but short, internal branches (see also Donoghue and Sanderson, 1992). Our approach to the twin problems of long-branch attraction and sampling error is to obtain relatively massive amounts of slowly evolving characters per taxon. This approach complements recent studies of basal angiosperm relationships that examined a greater number of taxa for fewer genes (Mathews and Donoghue, 1999; Qiu et al., 1999; Soltis, Soltis, and Chase, 1999; Parkinson, Adams, and Palmer, 1999). Fortunately, automated DNA sequencing technology now permits the rapid collection of a large number of characters. Primers are available for amplifying and sequencing three chloroplast genes that have been used successfully in studies of a broad array of angiosperm groups: *rbcL* (Zurawski, Clegg, and Brown, 1984); *ndhF* (Olmstead and Sweere, 1994; Kim and Jansen, 1995); and *atpB* (Hoot, Culham, and Crane, 1995). We describe here primers for 14 additional genes that are useful for amplifying and sequencing



Forward		Reverse	
		3R:	GATYGGAAATCCTGTATTTTMCA
1F:	CCMAAAAAACCMAACTCTGCCTT	5R:	GATAAGCCAATGATTTTTTTCC
2F:	ATCGTCAACAAGGGCGTTCTAGTG	6R:	CTATTTGCCTCTGCCATTCTATGAG
4F:	GAGATCCACCCTACAATATGGGG	7R:	CATTCTGGGAAAATSAAACTTCC
6F:	CTCATAGAATGGCAGAGGCAAATAG	8R:	GATAGAGGAATACATAGAGTTGAAC
7F:	GGAAGTTTSATTTTCCCAGAATG	B10R:	GAAAGCTTGAACCCAATTCCTAC
9F:	ATGGTTTCTCTTGGCTATATGG	10R:	CCTTCGTAGACGTCAGGAGTCCATTG
10F:	CAATGGACTCCTGACGTCTACGAAGG	C11R:	TTACACTCGTAGTCTCTGAAG
11F:	GAAGGATTCCTCGAAAAGTTAAGG	11R:	CCTTAACTTTTCGAGGAATCCTTC
12F:	CTTTCTGTTACTTCGAAAGTAGC	13R:	GCATACGTTTCATGCTTGTTTGAG
13F:	CTCAAACAAGCATGAAACGTATGC	14R:	GGTATAGTAGATGCTATCACACA

Fig. 1. Map of the primers used to amplify (open arrows) and sequence (all arrows except 10R) the chloroplast region that includes 3′ *rps12*, *rps7*, and *ndhB* and their introns. Primer 15R (5′-GAGATTTTGAGTCTCGCGTGTC; not shown on the map) is located in the *trnL* gene several 100 bp downstream of *ndhB*. The following primer pairs were typically used for PCR amplification of the region: 1F/7R; 6F/10R; 9F/13R; 11F/14R or 12F/15R. The scale is relative to *Nicotiana tabacum* sequence (GenBank accessions GBAN-Z00044 and GBAN-S54304) (primers are not drawn to scale). Degenerate sites follow IUPAC/IUB ambiguity codes. See the Appendix for more detailed primer information.

across the seed plants. These genes include three highly conserved introns and span five additional chloroplast regions.

The five additional regions examined include complete or partial coding sequences for genes from the Large Single Copy (LSC) and Inverted Repeat (IR) regions of the chloroplast genome (see Table 2). Of the 14 additional genes considered, ten are Photosystem II (*psb*) genes located in three distinct LSC regions. The remaining four are located in two IR regions in *Nicotiana*, *Oryza*, and *Zea*. They comprise three ribosomal protein genes and a gene for another NADH dehydrogenase (*ndh*) subunit. These genes were chosen to be at least as slowly evolving as those currently used to address basal angiosperm relationships. Genes in the IR have a six- to tenfold lower synonymous substitution rate than those in the single-copy regions (Wolfe, Li, and Sharp, 1987; Goremykin et al., 1996), and Photosystem II genes have some of the lowest synonymous substitution rates of single-copy chloroplast genes (Olmstead and Palmer, 1994). They have a correspondingly low level of multiple change per site. An important implication of using sequences with low synonymous substitution rates is that they will have low site-to-site heterogeneity in substitution rates, which enables a better fit to models of phylogeny reconstruction. We examined previously published GenBank sequences for these regions (results not shown) and found that they also have a very low observed frequency of multiple change, with the IR sequences in particular showing a low amount of repeated change at each variable site. We obtained new data from these regions for a broad range of exemplar taxa, using novel primers described here. We demonstrate the utility of these genes in sorting out basal angiosperm relationships, and also show that caution should be exercised over the finding (Parkinson, Adams, and Palmer, 1999; Qiu et al., 1999; Soltis, Soltis, and Chase, 1999) that the root node of the angiosperms has been definitively resolved (see also Graham et al., in press).

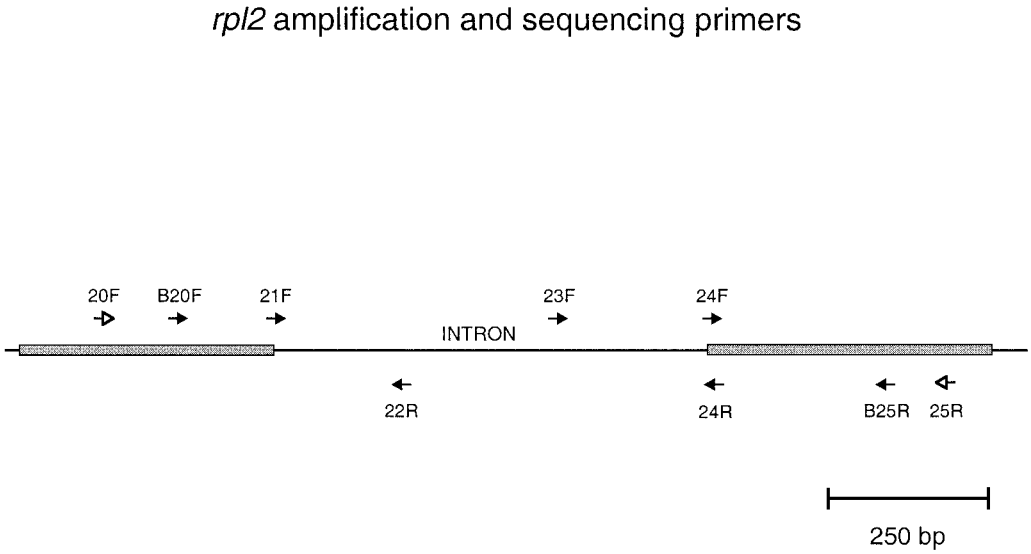
MATERIALS AND METHODS

Taxon sampling—Seed-plant taxa were sampled from a list of exemplar species compiled by the Green Plant Phylogeny Research Coordination Group

TABLE 2. Genes examined in the current study and their products. IR = Inverted Repeat; LSC = Large Single Copy region; SSC = Small Single Copy region.

Gene	Genomic location	Gene length (bp)[a]	Portion of gene examined[a]	Protein product
3′ *rps12*[b,c]	IR	794+	34–794 (761)	Ribosomal protein S12
rps7	IR	892	1–892 (892)	Ribosomal protein S7
ndhB[b]	IR	2212	1–2137 (2137)	NADH dehydrogenase subunit B
rpl2[b]	IR	1491	138–1407 (1270)	Ribosomal protein L2
atpB	LSC	1434	1–1497 (1497)	H^+-ATPase subunit β
rbcL	LSC	1497	27–1375 (1349)	Rubisco large subunit
ndhF	SSC	2223	29–1317 (1289)	NADH dehydrogenase subunit F
psbD	LSC	1062	81–1062 (982)	Photosystem II D2 protein
psbC[d]	LSC	1422	1–1255 (1255)	Photosystem II CP43 protein
psbE	LSC	252	22–252 (230)	Cytochrome b-559, α subunit
psbF	LSC	120	1–120 (120)	Cytochrome b-559, β subunit
psbL	LSC	117	1–117 (117)	Photosystem II subunit
psbJ	LSC	123	1–91 (91)	Photosystem II subunit
psbB	LSC	1527	30–1527 (1498)	Photosystem II CP47 protein
psbT	LSC	105	1–105 (105)	Photosystem II subunit
psbN	LSC	132	1–132 (132)	Photosystem II subunit
psbH	LSC	222	1–109 (109)	Photosystem II subunit

[a] Reference taxon = *Nicotiana tabacum*.
[b] Includes a single intron with lengths (relative to *Nicotiana*): 3′ *rps12* (536 bp); *ndhB* (679 bp); *rpl2* (666 bp).
[c] The region examined is part of a *trans*-spliced gene.
[d] 53-bp overlap with *psbD*.



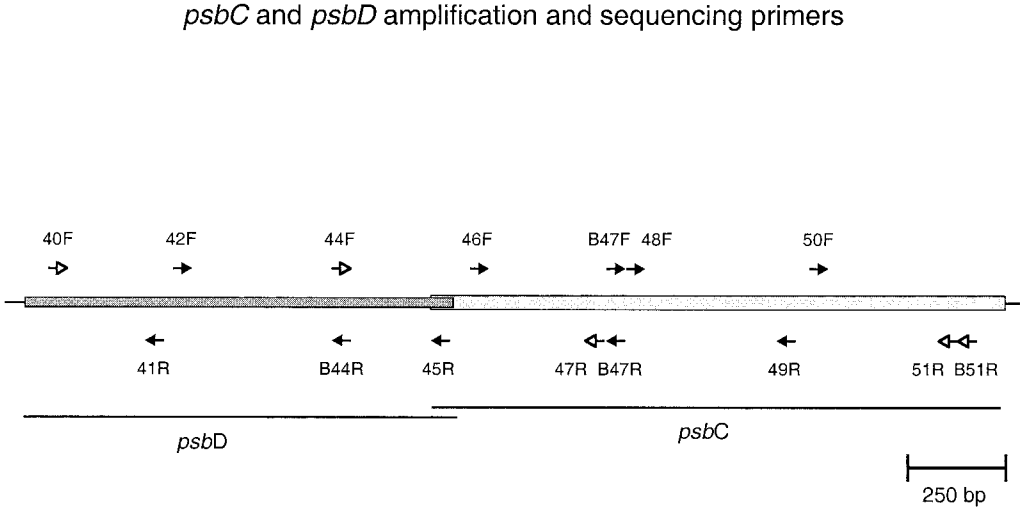
Forward

20F: AAAGGTCGTAATGCCAGAGGAAT
B20F: GTAACCATAGAATACGACCC
21F: CCCTACCTTTGAGTGCGGTTTGAA
23F: GCTACATGAAGAACATAAGCCAGATG
24F: CAACCRATATGCCCTTAGGCAC

Reverse

22R: GTCTTCTCCATATTACYATATCT
24R: GTGCCTAAGGGCATATYGGTTG
B25R: GGGTTCATARCTACTCCTCTTAC
25R: TTCCAAGYGCAGGATAACCCCA

Fig. 2. Map of primers used to amplify (open arrows) and sequence (all arrows) the chloroplast gene *rpl2* and its intron. Primer pair 20F/25R was typically used for PCR amplification. Primers B20F and B25R were often used as alternate sequencing primers to primers 20F and 25R, respectively. The scale is relative to *Nicotiana tabacum* sequence (primers are not drawn to scale).



Forward

40F: ATGACTGGTTACGRAGGGACCG
42F: TAATAGGTTTYATGYTACGTCAAT
44F: GGKGTTGCTTTTTCCAATAAACG
46F: GGTTTYGCTTGGTGGGCYGGGAATGC
B47F: GGAAAGATAGRAAYAAAATGAC
48F: TTAGGTATTCACYTAATYTTGTTAGG
50F: CTCAAGCATTTACTTTTYTAGTTAGAGA

Reverse

41R: ATTGASACCAACGAGTAAAATC
B44R: CGTTTATTGGAAAAAGCAACCCC
45R: CCTCCTCAGGGAATATAAGRTTTTCAT
47R: CATAACCRAAGAAKGGAAAAGATTC
B47R: GTCATTTTRTTYCTATCTTTCC
49R: GTATTATTGAACCAGACAAAACAACAAGC
51R: GCTAACCAACTTCTAGGRGARACAT
B51R: TACGAAGAAGAARAATCCTAGAAC

Fig. 3. Map of the primers used to amplify (open arrows) and sequence (all arrows except 47R) the overlapping chloroplast genes *psbD* and *psbC*. The following primer pairs were typically used for PCR amplification of the region: 40F/47R and 44F/51R or 44F/B51R. Primers B47F and 51R were often used as alternate sequencing primers to primers 48F and B51R, respectively. The scale is relative to *Nicotiana tabacum* sequence (primers are not drawn to scale).

(GPPRCG), a DOE/NSF/USDA-funded program designed to coordinate and stimulate research on green plant phylogeny. An initial subset of 16 exemplar angiosperms was chosen to represent most of the major basal lineages suggested by other molecular and morphological studies (Table 1). We concentrated on taxa with ''Priority 1'' status on the GPPRCG list. Complete chloroplast genome sequences are already available for three of these taxa (*Nicotiana tabacum*, *Oryza sativa*, and *Zea mays*) and one outgroup taxon (*Pinus thunbergii*). With the inclusion of *Gnetum gnemon* and *Ginkgo biloba* as outgroup taxa, all extant seed-plant groups except Cycadales were represented in the analysis.

Primer design—The new primers permit amplification and sequencing of the following genes from five regions of the chloroplast genome (Table 2; Figs. 1–5): (1) the *rps12* 3′-intron and exons (referred to as the ''3′ *rps12*'' gene here), *rps7*, and *ndhB*; (2) *rpl2*; (3) *psbD* and *psbC*; (4) *psbE*, *psbF*, *psbL*, and *psbJ*; and (5) *psbB*, *psbT*, *psbN*, and *psbH*.

The first two regions are located in the Inverted Repeat and include three ribosomal protein genes (*rpl2*, 3′ *rps12*, and *rps7*) and an NADH dehydrogenase subunit gene (*ndhB*). The 3′ *rps12*, *rps7*, and *ndhB* cluster of genes (Fig. 1) lie several kilobases away from the *rpl2* gene (Fig. 2). The typical angiosperm IR includes all four genes (e.g., *Nicotiana*, *Oryza*, and *Zea*). They lie outside the IR of *Marchantia polymorpha* and the highly reduced IR of *Pinus thunbergii*. The 3′ *rps12* gene includes the last two of three exons of this divided, *trans*-spliced gene (Zaita et al., 1987). The 5′ exon of this gene (not sampled here) is located in the LSC region. A single Group II intron is present in each of the 3′ *rps12*, *ndhB*, and *rpl2* genes, and small intergenic spacers (IGSs) separate 3′ *rps12*, *rps7*, and *ndhB* (Fig. 1).

The three *psb*-gene containing regions we developed primers for (Figs. 3–5) are all located in the LSC region of the genome. Seven of the *psb* genes are quite small, ~100–200 bp in length. The genes within each region are coordinately transcribed, although *psbB*, *psbT*, and *psbH* are part of a larger transcription unit, as are *psbD* and *psbC* (see Gruissem and Tonkyn, 1993). The *psbN* locus lies between the *psbT* and *psbH* genes (Fig. 5), but is transcribed from the opposite strand. The *psbC* gene partly overlaps *psbD* (Fig. 3) and is cotranscribed with it, but also transcribed from a separate point within *psbD* (Yao et al., 1989). The *psbD* and *psbC* reading frames are not in phase with each other. Small IGSs separate the other *psb* genes (Figs. 4–5). Alternative promoters have also been found within the *psbE-psbF-psbL-psbJ* operon (Haley and Bogorad, 1990).

Primer sites were spaced at most ~450–530 bases apart on each strand, a distance chosen because it allows a reasonable overlap using an ABI Prism 377 automated sequencer (PE Biosystems, Foster City, California, USA). Staggered spacing of primers on the forward and reverse strands minimizes the danger of sequence signal expiring at the same point on different strands. The IGS regions were rejected as sites for primer design, but several primers were placed inside the three introns or at intron/exon boundaries, because these sequences were found to be slowly evolving. These primers (Figs. 1–5; Appendix) should fail in taxa lacking introns, but that did not apply to any taxa examined in this study.

Choosing primer sites with as little variation as possible across the aligned

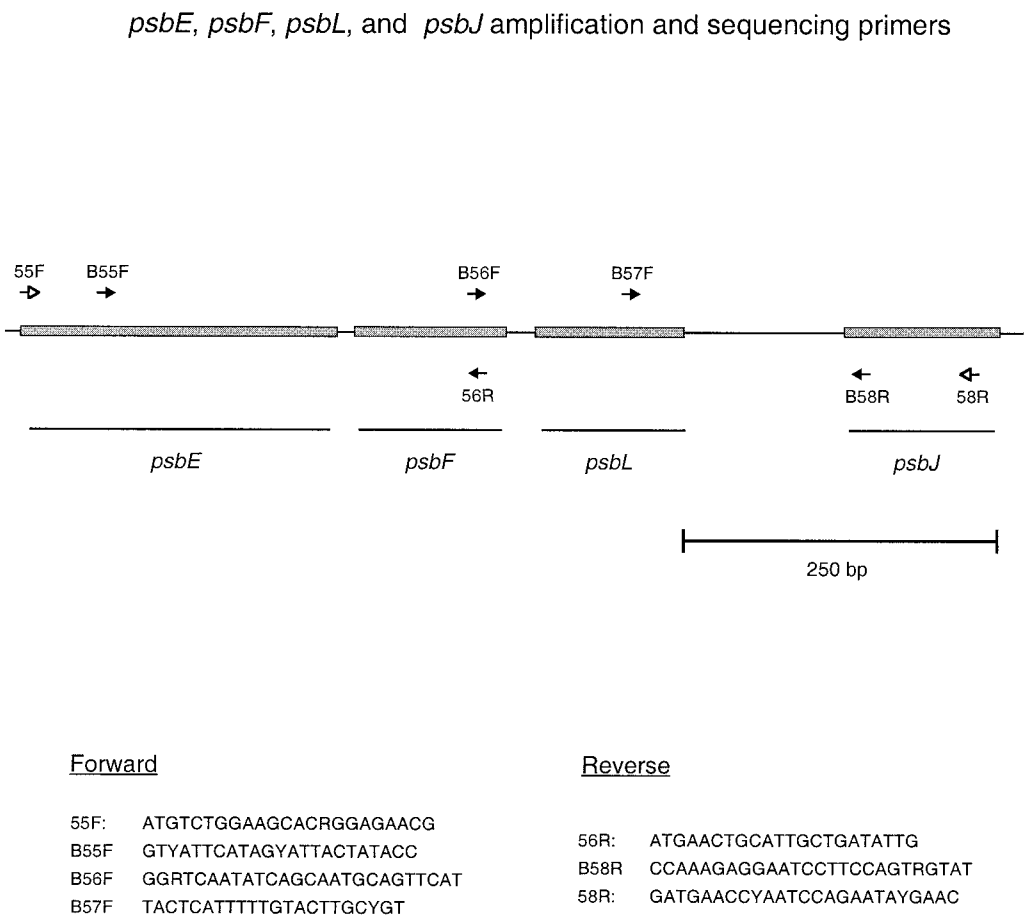


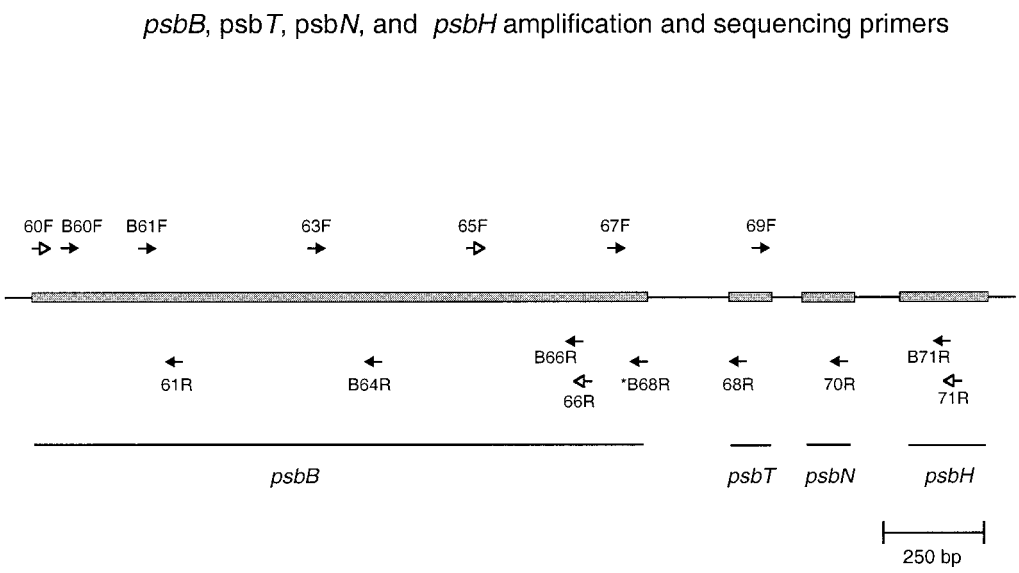
Fig. 4. Map of the primers used to amplify (open arrows) and sequence (all arrows) the chloroplast genes *psbE, psbF, psbL,* and *psbJ*. Primer pair 55F/58R was typically used for PCR amplification. Primers B55F and B57F and B58R were often used as alternate sequencing primers to primers 55F, B56F and 58R, respectively. The scale is relative to *Nicotiana tabacum* sequence (primers are not drawn to scale).

taxa was made easier by the generally low substitution rates in the chloroplast genome, particularly for the IR sequences. Sequences of *Marchantia polymorpha*, *Pinus thunbergii*, and an assortment of angiosperm taxa available in GenBank were considered for primer design in the different regions. Alignments for each region were obtained using Clustal W (Thompson, Higgins, and Gibson, 1994). For primers in protein-coding sequences, the 3′-most base (the one leading into the sequence) was generally chosen to be a second-codon position, but occasionally a first- or third-codon position was used, the latter only if it belonged to a conserved codon for a twofold or nondegenerate amino acid. Variable sites within the primers regions were accounted for using partial nucleotide degeneracy in the primer sequence. Primers were assessed for duplex formation in Amplify 1.2 (Engels, 1993) and were discarded if they had a T_m lower than ~60°C or obvious hairpin regions. All primers were at least 20 bases long to maximize the T_m and, hence, the specificity of binding. Whenever possible, each primer was positioned so that it ended in a one or two base CG-clamp (G and C are the strong-pairing bases). Replacement or alternate primers were necessary in a few cases (primers with B and C prefixes in Figs. 1–5) where amplifications and cycle-sequencing reactions failed to work for all species.

Amplification and sequencing protocols—The following thermocycler profile was used for Polymerase Chain Reaction (PCR) amplifications: (1) initial denaturing at 94°C for 5 min; (2) 30 cycles of the following: denaturation at 94°C for 1 min, annealing at 45°C for 1 min, extension at 72°C for 2 min; (3) final extension at 72°C for 15 min. The reactions were performed in 50-μL volumes, using 25 pm of each primer. QIAquick PCR purification columns (QIAgen Inc., Valencia, California, USA) were used to purify PCR products following manufacturer's instructions, except that 30 μL of water preheated to 60°C was used for elution.

The primer pairs typically used in amplification are noted in the Appendix (see also Figs. 1–5), together with the sequencing primers (and alternates) used for each fragment. The following alternative amplification primer-pairs were occasionally used when the PCR products were absent or weak: region 1: 9F/14R instead of 9F/13R, and 12F/15R instead of 11F/14R; region 2: 20F/24R plus 21F/25R instead of 20F/25F; region 3: 40F/B47R instead of 40F/47R, 44F/51R instead of 44F/B51R; region 4: B55F/58R or 55F/B58R instead of 55F/58R. The *atpB* gene was sequenced for a few taxa where these were not already available, using the primers designed by Hoot, Culham, and Crane (1995). We sequenced *ndhF* using primers designed by Olmstead and Sweere (1994) and Kim and Jansen (1995). Only the 5′ end of *ndhF* (bases 29–1317 in tobacco) was included, because the 3′ end of the gene exhibits extensive length variation, in combination with an increased substitution rate (Olmstead and Sweere, 1994; Kim and Jansen, 1995; Olmstead and Reeves, 1995).

An ABI Prism dRhodamine terminator cycle sequencing ready reaction kit (PE Applied Biosystems) was used to set up sequencing reactions following the manufacturer's instructions, except that 35 ng of PCR product were used per half-reaction. For cycle sequencing, 25 cycles of the following conditions were used: (1) denaturation at 96°C for 10 s; (2) annealing at 45°C for 5 s; (3) extension at 60°C for 4 min. Individual cycle sequencing products were cleaned on a Sephadex column and precipitated using an unheated vacuum centrifuge for half an hour. Resuspended products were run on an ABI Prism 377 automated sequencer. All regions were sequenced at least twice for each taxon, and with a few minor exceptions these represent both forward and reverse strands. Because a large number of PCR and sequencing products



Forward		Reverse	
60F:	ATGGGTTTGCCTTGGTATCGTGTTCATAC	61R:	TCCCAATAYACCCAATGCCAGATAG
B60F:	CATACAGCTCTAGTTKCTGGTTGG	B64R:	CTTGCTGRAAGTATCCYTGATCCC
B61F:	CGGGTMTTTGGAGTTAYGARGG	B66R:	CCCCTTGGACTRCTACGAAAAACACC
63F:	GGATTRCGTATGGGMAATATTGAAAC	66R:	CCAAAAGTRAACCAACCCCTTGGAC
65F:	TGCCTACTTTTTTTGAAACATTTCC	*B68R:	GTAGTTGGATCTCCAAGTTTTTGG
67F:	GAGATGTTTTTGCTGGTATTGA	68R:	AAYGTATAAACCAATGCTTCCAT
69F:	TCGCTATCTTYTTYCGAGAACCRC	70R:	TATCTGGTTTACTTGTAAGYTTTACTGG
		B71R:	CCAGGAGCTACTTTACCATATTC
		71R:	CCCATMAAAGGAGTAGTYCCCC

Fig. 5. Map of the primers used to amplify (open arrows) and sequence (all arrows except 66R) the chloroplast genes *psbB, psbT, psbN,* and *psb*H. The following primer pairs were typically used for PCR amplification of the region: 60F7/66R and 65F/71R. Primers B60F and B61F were often used as alternate sequencing primers to primer 60F; primers *B68R and B71R were often used as alternate sequencing primers to 68R and 71R, respectively. Note that the gene *psbN* is on the strand opposite to the co-ordinately transcribed *psbB, psbT,* and *psbH* genes. The scale is relative to *Nicotiana tabacum* sequence (primers are not drawn to scale).

were handled, a control for sample provenance was included by obtaining partial sequence from at least one replicate PCR product (from amplification reactions performed on different days) for each major region and taxon.

Data compilation—Sequencher 3.0 (Gene Codes Corp., Ann Arbor, Michigan, USA) was used to compile contiguous sequences from electropherograms generated on the automated sequencer. PCR primer sequences incorporated into sequenced products were excluded from contigs. Completed contigs were exported from Sequencher as text files and aligned across taxa using Clustal W (Thompson, Higgins, and Gibson, 1994). Alignments exported in PHYLIP format (Felsenstein, 1995) were then imported into Se-Al 1.0 (Rambaut, 1998) for minor manual adjustment. Final alignments were then exported in PHYLIP noninterleaved format and imported into PAUP* 4.0 beta (Swofford, 1999). This process was repeated for each region. Finally, a master file containing alignments for each region was assembled (with FORMAT set to ''interleaved''). We were careful to maintain taxon order among noninterleaved alignments for the different regions, as taxon labels in interleaved files were not verified by the version of PAUP* we used. Gaps were treated as missing data in the analysis but retained as distinct records. Alignments are available on request from SWG.

Coordinates for gene, exon, and intron borders were determined by comparison with tobacco sequences, and this information was used to derive CHARSETs (character sets) in PAUP* to define the nucleotides used in each analysis. Finally, a separate binary character matrix (Table 3) was assembled for indel (insertion/deletion) characters in the intron and coding regions. This matrix was appended to the other characters in the master matrix. We excluded indels in IGS sequences from consideration because it was harder to determine their homology than in intron and coding sequences. Indels were also ignored if their homology could not be inferred unambiguously (in general this was only a problem for the outgroup seed plant taxa) or if they represented unique events in the outgroups.

GenBank accession numbers for new sequences are as follows (the prefix GBAN- has been added to link the online version of *American Journal of Botany* to GenBank, but is not part of the actual accession number): (1) GBAN-AF123771–AF123784 for the *rps12, rps7,* and *ndhB* region; (2) GBAN-AF123785–AF123798 for *rpl2*; (3) GBAN-AF123813–AF123827 for the *psbD* and *psbC* region; (4) GBAN-AF123828–AF123842 for the *psbE, psbF, psbL,* and *psbJ* region; (5) GBAN-AF123843–AF123857 for the *psbB, psbT, psbN,* and *psbH* region; (6) GBAN-AF123799–AF123812 for *ndhF*; (7) GBAN-AF187058–AF187061 for *atpB.* The new *Amborella trichopoda* sequences for these seven regions are GBAN-AF235041–AF235047. Source information for these taxa and GenBank numbers for previously published chloroplast sequences used in this study are given in Table 1.

Phylogenetic analysis—Only coding and intron sequences and indel char-

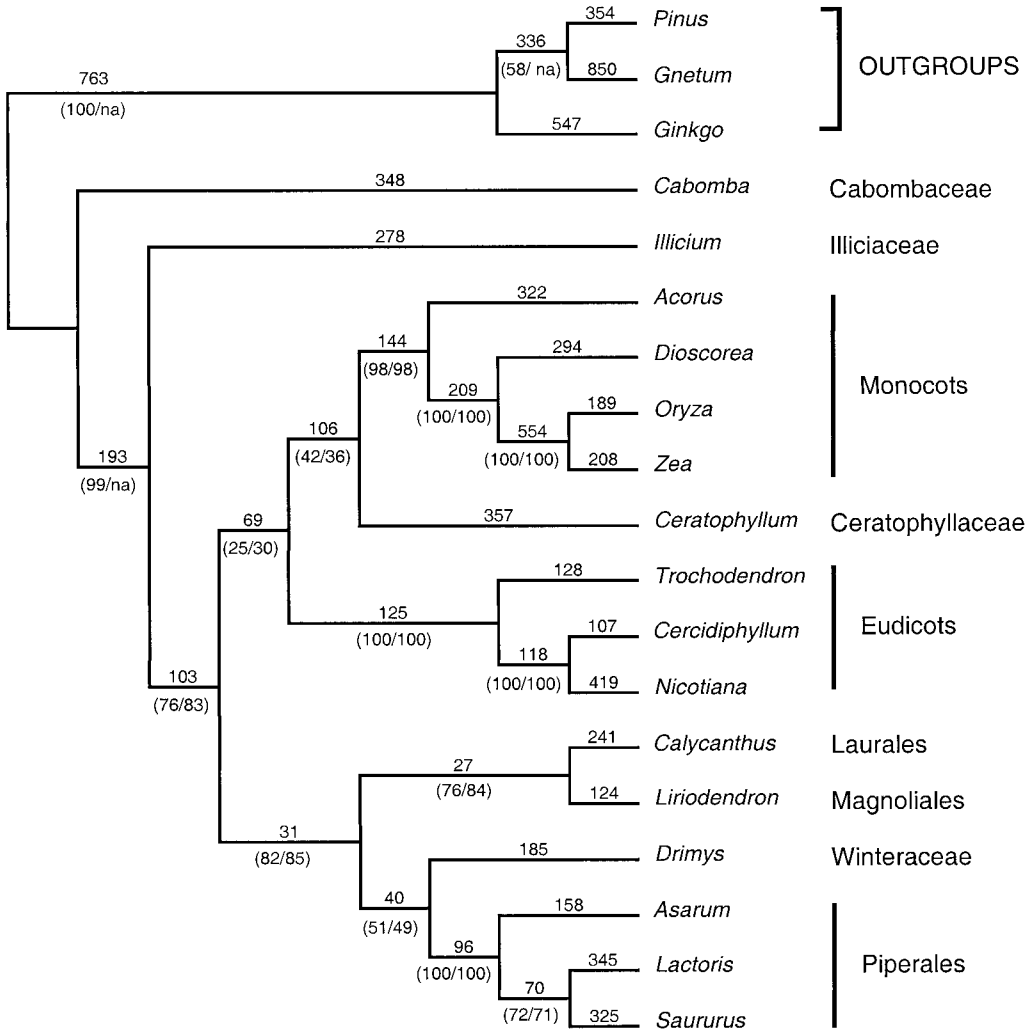


Fig. 6. Single most parsimonious tree found using combined coding sequence, intron, and indel data from 17 chloroplast genes. The same ingroup topology was found using only angiosperm taxa. Taxon names follow APG (1998). Branch lengths are indicated above branches (computed using ACCTRAN optimization). Numbers below branches are the percentage of bootstrap replicates supporting that branch in analyses that include and exclude the outgroups respectively. See Table 5 for tree length and fit statistics.

acters were considered. The IGS sequences were excluded from phylogenetic analysis because it was often difficult to determine nucleotide homology. Heuristic searches were performed in PAUP* with all characters and character-state changes equally weighted. MULPARS and ''Steepest Descent'' options were activated, and 100 random addition replicates were performed for each search. All the data were analyzed simultaneously and also in the following data partitions: the IR data (coding and introns) combined; the Photosystem II genes combined; and the three remaining single-copy genes combined (*atpB* plus *ndhF* plus *rbcL*). These roughly equally sized data partitions were considered because they group the underlying data into at least two fairly natural groups: the IR sequences collectively evolve at a slower rate than the other sequences, and the *psb* genes all code for proteins in the Photosystem II complex. The appropriate indels were included for each data partition. All analyses were repeated with and without outgroup taxa, because these were by far the longest branches on the trees. The angiosperm subtree found from the most parsimonious tree from the combined analysis (see Results) was used to estimate the gamma-distribution shape parameter, alpha, available under the ''Tree Scores/Likelihood'' option in PAUP*.

Bootstrap analysis (Felsenstein, 1985) was performed with the same search criteria, except that one random-order entry starting tree was used for each of the 100 bootstrap replicates. Bootstrap analysis provides biased, but usually conservative, estimates of the accuracy of individual clades (Hillis and Bull, 1993). Hillis and Bull showed that branches supported by ~ 70% or more replicates tend to be representative of the true phylogeny so long as rates of change are not very high or very unequal among lineages (see Felsenstein and Kishino, 1993, for a slightly different interpretation). We refer to branches with at least this much support as ''well-supported'' while recognizing that phenomena such as long-branch attraction can also lead to erroneously high support values. Bootstrap analyses were repeated on the three data partitions. The mean bootstrap support from each of these data partitions also was determined for the angiosperm subtree inferred using all the data (i.e., the average bootstrap support from each data partition for the 13 nonterminal branches in this tree). The incongruence length difference (ILD) test of Farris et al. (1994) was used to assess the significance of incongruence among these subsets of the available chloroplast data. The same heuristic search criteria were used as above, except that ten random addition replicates were used for each of the 100 permutation replicates.

RESULTS

The *rpl2* intron is known to be missing in several eudicot groups (Downie et al., 1991), but is present in *Pinus*, *Ginkgo*,

TABLE 3. Indel events in coding and intron sequences (IGS sequences and unique or ambiguous indels in outgroup taxa ignored). These inferences are restricted to the 19-taxon study. Taxa with state = ''1'' have extra bases relative to those with state ''0.''

Label (Fig. 7)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Gene							3′ *rps12*							*rps7*							*ndhB*						
Fragment							Intron					Exon 2			Exon 1			Intron									
Location[a] (bp)	43	158	173	185	225	306	309	383	382	431	445	23	23	239	57	32	42	58	112	123	157	192	222	269	279	320	321
Size	3	5	6	1	5	3	5	1	10	8	5	3	5[c]	3	9	8	5	1	6	21	11	3	10	2	5	1	4
Inferred polarity of indel event(s)[b]	D?	ID	I	D	ID	D	I	ID	D[d]	ID	I	(I	I)[d]	I	I	D	I	D	I	D	D	I	D	I	I	D	I
Pinus	1	*	0	1	0	1	0	1	1	1	0	*	*	0	?	?	?	?	?	?	?	?	?	?	?	?	?
Gnetum	0	*	*	1	0	1	0	1	1	1	0	*	*	0	?	?	?	?	?	?	?	?	?	?	?	?	?
Ginkgo	1	0	0	1	1	1	*	1[e]	1	1	*	0	0	0	0	1[e]	*	1	0	1[e]	1	0	1	0	0	1	*
Cabomba	1	0	1	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0	1	1	0	1	0	1	1	0
Illicium	0	1	0	1	1	1	0	*	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	0	0	1	1
Acorus	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Dioscorea	0	1	0	1	1	0	0	0	1	0	1	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Oryza	0	1	0	0	1	1	1	0	1	0	0	1	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0
Zea	0	1	0	0	1	1	1	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	1	0
Ceratophyllum	0	1	0	1	1	1	0	0	1	0	0	0	1	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Trochodendron	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Cercidiphyllum	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	1	1	0	0	1	0
Nicotiana	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Calycanthus	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Liriodendron	0	1	0	1	1	1	1	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Drimys	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0
Asarum	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	1	1	1	0	1	0	0	1	0
Lactoris	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0
Saururus	0	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	0	0	1	0

[a] Corresponds to first base preceding the indel in the gene or gene-fragment (relative to *Nicotiana tabacum*).

[b] I = Insertion; D = Deletion; ID = Polarity of indel event ambiguous.

[c] Reading frame is disrupted: a novel stop codon is present in *Ceratophyllum* three codons downstream of the insertion point, the penultimate codon in *Nicotiana*.

[d] Underline: independent but potentially phylogenetically nested indel events in the same location. Parenthesis: independent and non-nested indel events in the same location.

[e] Indicates additional ambiguous (or autapomorphic) indel events are found in the same location in outgroup taxa (not listed here).

[f] Truncation of protein product is (*Zea*) or may be (*Nicotiana*; ignored in analysis) associated with one or more indel events that extend across the coding and intergenic regions.

* Data present but indel homology to other taxa uncertain; coded as ''missing data'' in the analysis. ? = missing data sequence.

and all the angiosperms considered here. Several genes for particular taxa had to be omitted because they were missing, nonamplifiable, or could not be sequenced adequately. These characters were coded as missing data in the phylogenetic analyses. They are (by taxon): (1) *ndhF* and *ndhB* in *Pinus* (an *ndh*B pseudogene is present but ignored here); (2) *ndhF*, *ndhB*, and *rpl2* in *Gnetum* (none of which could be amplified successfully); (3) *psbJ* in *Acorus*; (4) *psbH* in *Asarum.* The last two genes could not be sequenced because of homopolymer stretches in the IGS regions preceding each gene. Fairly substantial parts of several genes were excluded due to difficulty in obtaining amplifications or high-quality sequence: *Cabomba atpB* (213 bp missing from the 5′ end relative to the sequences); *Dioscorea ndhF* (261 bp from the 5′ end); *Ginkgo ndhF* (222 bp from the 5′ end). Several GenBank *rbcL* sequences were also substantially foreshortened: *Cabomba* (192 bp missing at the 3′ end) and *Acorus* (138 bp missing at the 3′ end).

A few genes also showed minor truncation or expansion of their inferred reading frame. Sequence beyond the truncation/expansion point was excluded in phylogenetic analysis, unless nucleotide homology could be assigned unambiguously. For example, the start codon inferred in published angiosperm *ndhB* sequences is upstream from that in *Marchantia* (GenBank accession GBAN-X04465) and *Ginkgo,* such that the 5′ exon of *ndhB* is 17–18 codons longer in the flowering plants. Because sequence upstream of the start codon in *Ginkgo* can be aligned unambiguously with the corresponding coding sequence in the angiosperms, this noncoding sequence was included in analyses.

RNA edit sites are known or suspected in several chloroplast genes (e.g., Maier et al., 1995; Freyer, Kiefer-Meyer, and Kössel, 1997). Apart from *Pinus, Gnetum, Ginkgo,* and the four monocot taxa, a previously reported edit site within the initiation codon of *psbL* (Kudla et al., 1992; Bock et al., 1993) is inferred in all the taxa we examined here, including *Amborella.* An edit from C to U is necessary to produce a functional translation initiation codon for this gene in these taxa. Because all the sequences we considered were derived from DNA, the RNA edit sites in this and other genes should not have an intrinsically misleading effect on the phylogenetic analysis we performed (Bowe and dePamphilis, 1996).

In most cases the lengths of noncoding regions (introns and IGSs) did not vary greatly across the taxa, either within the angiosperms or across the seed-plant taxa (Table 4). The greatest range of length variation was in the IGS region between the *psbB* and *psbT* loci (Table 4). In all cases the standard deviation of length variation across examined sequences was <30 bp. This amount of length variation was thus not large enough to interfere with the ability to generate overlapping sequencing fragments for any region or taxon examined here.

Phylogenetic analysis—Two tiers of phylogenetic analyses were performed. The major set of analyses focused on a core

TABLE 3. Continued.

Label (Fig. 7)	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	n/a[f]	53	54
Gene	*ndhB*				*rp12*														*rp12*					*ndhF*	*psbT*		*psbH*	*atpB*
Fragment	Intron				Intron														Intron									
Location[a] (bp)	510	521	584	594	213	234	243	276	325	326	331	341	346	347	383	452	486	510	510	575	639	650	650	647	102	105	32	157
Size	5	12	6	2	4	4	10	9	3	11	1	1	1	34-35	5	1	21	5	2	3	2	1	4	3	6+	3?	3	6
Inferred polarity of indel event(s)[b]	I	I	I	I	I	I	D	I	D	ID[d]	D	D	D	I[e]	I	I	I	(I	I)[d]	D	D	D	I	I	(D	ID)[d,f]	I	D
Pinus	?	?	?	?	1	1	*	0	1	1	1	1	1	1	0	0	0	*	*	*	1	1	0[e]	?	1	1	0	0
Gnetum	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	1	1	0	0
Ginkgo	0	0	0	0	0	1	1[e]	0	1	1	1	1	1	1	0	0	0	*	*	1	1	1	0	0	1	1	0	0
Cabomba	1	0	0	0	0	0	1	0	1	0	1	1	1	0	1	0	0	0	0	1	1	1	0	1	1	1	0	0
Illicium	0	0	1	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0
Acorus	0	0	0	0	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0
Dioscorea	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	0	1	1	1	1	0	1	1	0	1
Oryza	0	1	0	0	0	0	1	0	0	*	0	1	0	0	0	0	0	1	0	1	0	1	0	0	1	1	0	0
Zea	0	1	0	0	0	0	1	0	0	*	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0
Ceratophyllum	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0
Trochodendron	0	0	0	0	0	0	1	1	1	0	1	1	1	0	0	0	0	0	0	1	1	1	0	1	1	1	0	0
Cercidiphyllum	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0
Nicotiana	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0
Calycanthus	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0
Liriodendron	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0	1	1	0	0
Drimys	0	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0
Asarum	0	0	0	1	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0
Lactoris	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	1	1	0	0	1	1	1	1	0	1	1	0	0
Saururus	0	0	0	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	1	1	0	0

of 19 taxa (Tables 3–5; Figs. 6–8). A second set of analyses were performed that included one or two additional taxa (*Amborella trichopoda* alone or *Amborella* and *Nymphaea odorata* together; Fig. 9, and see below).

Nineteen-taxon analyses—The same angiosperm subtree was seen in the single most-parsimonious trees inferred using the combined data from analyses that included (Fig. 6) or excluded the three outgroup taxa (Table 5, topological distance of zero). The combined data strongly supported the water lilies, represented by *Cabomba*, as the sister group to the rest of the angiosperms (Fig. 6). *Illicium* was resolved as the sister group to the remaining angiosperms, a relationship that was also well supported by bootstrap analysis. Four other major groups were well supported as monophyletic groups, at the current taxon sampling. These are: the monocots (including *Acorus*), the eudicots, Piperales (sensu APG, 1998; includes Piperaceae, Aristolochiaceae, and *Lactoris*), and a group composed of three woody magnoliids representing Magnoliales (*Liriodendron*), Winteraceae (*Drimys*), and Laurales (*Calycanthus*), together with Piperales. Relationships among these clades (and *Ceratophyllum*) were not strongly supported by bootstrap analysis (Fig. 6), but in the most parsimonious tree *Ceratophyllum* was the sister group of the monocots, and the eudicots were the sister group of that clade. The mean bootstrap support for the 13 nonterminal branches in the unrooted tree of 16 angiosperm species was nearly 80%.

Angiosperm monophyly was supported by 100% of replicates for all three data partitions. Five uncontradicted indels also support the monophyly of the angiosperms (Fig. 7), presuming that the root of the seed plants falls outside the flowering plants. In most cases the indel characters were not synapomorphies for any of the phylogenetic structure inferred within the angiosperms, at least at this level of taxon sampling. However, one indel supports *Dioscorea* plus the two grasses, and ten indels support the monophyly of the two grass taxa sampled. Two indels also support the basal position of *Cabomba*, one of which is apparently a convergence with an indel in *Gnetum* (Table 3; Fig. 7). A single-base indel (inferred to be a deletion; Table 3) seems to link Winterales and Magnoliales, but this is not inferred to be homologous between these two taxa on the most parsimonious tree (Fig. 6) since the same base is not lacking in *Calycanthus* or Piperales (Table 3; Fig. 7). A medium-sized inversion located between *rps7* and *ndhB*, which likely evolved in parallel in *Calycanthus* and the water lilies, has been noted previously (Graham and Olmstead, 2000).

Of three distinct data partitions considered here (the IR genes, the Photosystem II genes, and the three other single-copy genes), the Photosystem II data had the highest mean

TABLE 4. Variation in lengths of noncoding regions. These statistics are derived from the 19-taxon study.

Noncoding region	Mean length[a] (mean in angiosperms)	Standard deviation[a] (SD in angiosperms)	Range[a] (range in angiosperms)
3′ *rps12* intron	536 (537)	15 (3)	487–575 (526–541)
3′ *rps12* - *rps7* IGS	56 (57)	4 (3)	48–60 (52–60)
rps7-ndhB IGs	313 (313)	12 (12)	278–326 (278–326)
ndhB intron	701 (699)	14 (11)	679–736 (679–715)
rp12 intron	669 (666)	14 (8)	654–716 (654–690)
psbE-psbF IGS	9 (10)	1 (1)	9–14 (9–14)
psbF-psbL IGS	24 (24)	5 (4)	18–38 (21–38)
psbL-psbJ IGS	126 (128)	11 (11)	113–161 (116–161)
psbB-psbT IGS	175 (182)	26 (12)	79–200 (162–200)
psbT-psbN IGS	63 (60)	10 (6)	48–86 (48–73)
psbN-psbH IGS	104 (108)	11 (7)	79–128 (102–128)

[a] To nearest basepair, across taxa in the study (where known).

TABLE 5. Summary statistics for phylogenetic trees of the basal angiosperms and outgroup taxa from various combined analyses. These statistics are derived from the 19-taxon study. (IR = Inverted Repeat; CI = consistency index; RI = retention index; MP = maximum parsimony; indel = insertion/deletion.)

Data set analyzed[a]	No. characters (no. informative)	No. of MP trees/Length	CI	RI	Distance[b]	Mean bootstrap[c]	Alpha[d]
Combined IR							
All 19 taxa	5056 (525)	12/1678	0.810	0.649	(4-10)		
Angiosperms only	5056 (283)	27/889	0.795	0.626	2-10	61%	0.275/0.268/0.090
rbcL, *atpB*, and *ndhF* combined							
All 19 taxa	4165 (969)	1/3700	0.570	0.433	(12)		
Angiosperms only	4165 (802)	1/2841	0.612	0.449	12	57.5%	0.484/0.375/0.253
Combined Photosystem II genes							
All 19 taxa	4585 (912)	1/3338	0.578	0.462	(4)		
Angiosperms only	4585 (621)	1/2162	0.607	0.434	4	68.4%	0.333/0.241/0.121
All data							
All 19 taxa	13806 (2406)	1/8763	0.616	0.469	(0)		
Angiosperms only	13806 (1706)	1/5913	0.636	0.459	n/a	79.7%	0.303/0.243/0.130

[a] Including genes, introns, and indels.

[b] Pairwise topological distance(s) in symmetric difference units (range where more than one tree) of MP tree(s) to the single MP tree found for the 16 angiosperm taxa using all the data. Values in parentheses are with respect to the same angiosperm subtree.

[c] Mean bootstrap support from data set for the 13 internal branches (taxon bipartitions) seen in the single tree found for the angiosperms using all the data.

[d] Parsimony-based estimate of the gamma shape parameter (alpha) for among-site rate heterogeneity, as determined by (respectively) the method-of-moments, the Sullivan et al. method, and the Yang-Kumar method (see Swofford, 1999) on the single MP tree found for the angiosperms using all the data (Fig. 6; angiosperm subtree).

bootstrap support for the angiosperm subtree inferred from all the data. The combined *atpB* + *ndhF* + *rbcL* data had the lowest mean bootstrap support for this tree, and the most parsimonious trees found with this data partition were also the most distinctive in shape (12 symmetric difference units compared to the angiosperm subtree inferred from all of the data). *Cabomba* was also resolved as the basal lineage in the most parsimonious tree(s) for all three chloroplast data partitions. For the IR and Photosystem II data partitions 100% of bootstrap replicates supported the *Cabomba* rooting, but for the combined *atpB* + *ndhF* + *rbcL* data, bootstrap support was almost equally split between a *Cabomba* and a *Ceratophyllum* rooting (44% of bootstrap replicates support the former, 49% the latter).

The IR data provided good support for most of the structure on the angiosperm subtree inferred from all the data, despite having fewer informative characters than either single-copy data partition. The mean support for the subtree from this data partition was 61% (Table 5). The IR data found more than one most parsimonious tree, but one of these was the closest in shape to the tree inferred from all of the data combined (two symmetric difference units; Table 5) of any tree found by the three data partitions considered. As might be expected for such slowly evolving characters, the IR data had by far the lowest amount of homoplasy as measured by two different homoplasy estimators, the consistency index, CI, and the retention index, RI (Table 5).

All of the genes were inferred to be very slowly evolving. No change was inferred at a majority of characters (78%) across the angiosperm tree derived from all the data (88, 68, and 77% of all characters were invariant for the combined IR data, the combined *atpB* + *ndhF* + *rbcL* data, and the combined *psb* data, respectively). The IR data had the lowest parsimony-based estimate of the gamma-distribution shape parameter alpha (Table 5). This is in part a function of its high number of apparently invariant characters, but also because three-quarters of all variable characters were inferred to change only once for this data class. A narrow majority of variable characters in the other two data partitions changed more than once (Fig. 8).

The trees inferred from all the data resolve *Lactoris* as the sister group of *Saururus.* This clade is also strongly supported by the Photosystem II data (89% bootstrap support). In contrast, the combined *atpB* + *ndhF* + *rbcL* data partition resolves a different clade, *Lactoris* plus *Asarum*, with 74% bootstrap support. Only 23% of bootstrap replicates support (*Lactoris*, *Saururus*) for this data partition. This partly accounts for the lower mean bootstrap support for the angiosperm subtree by this data partition. No local relationship of *Lactoris* was supported by $>$ 50% of bootstrap replicates for the IR data partition.

Two of the data sets also provided moderate to strong support for a taxon bipartition (branch) that rejects Gnetales as the sister group of the angiosperms. A (*Gnetum, Pinus*) taxon bipartition (Fig. 6) was supported by 58% of bootstrap replicates for the IR data, and 100% for the combined *atpB* + *ndhF* + *rbcL* data. This is in line with several recent molecular studies of the nuclear gene *RPB2* (A. Denton and B. Hall; personal communication), 18S rRNA (Chaw et al., 1997) and the five-gene, three-genome study of Qiu et al. (1999). In contrast, the Photosystem II data support a (*Gnetum,* angiosperms) bipartition in 99% of bootstrap replicates, a result in line with the recent three-gene study of Soltis, Soltis, and Chase (1999).

The strong disagreement concerning outgroup relationships and the more modest one involving the local position of *Lactoris* are the only cases where conflicting tree structure among data partitions was well supported by bootstrap analysis. When the ILD test of Farris et al. (1994) was performed on the three data partitions, significant or nearly significant heterogeneity was found with the outgroup taxa included or excluded (P = 0.01 and 0.07, respectively). By excluding *Lactoris* and the three outgroup taxa, no significant heterogeneity was indicated (P = 0.26).

Analyses with Amborella included—When *Amborella trichopoda* was added to the core data set, it was strongly

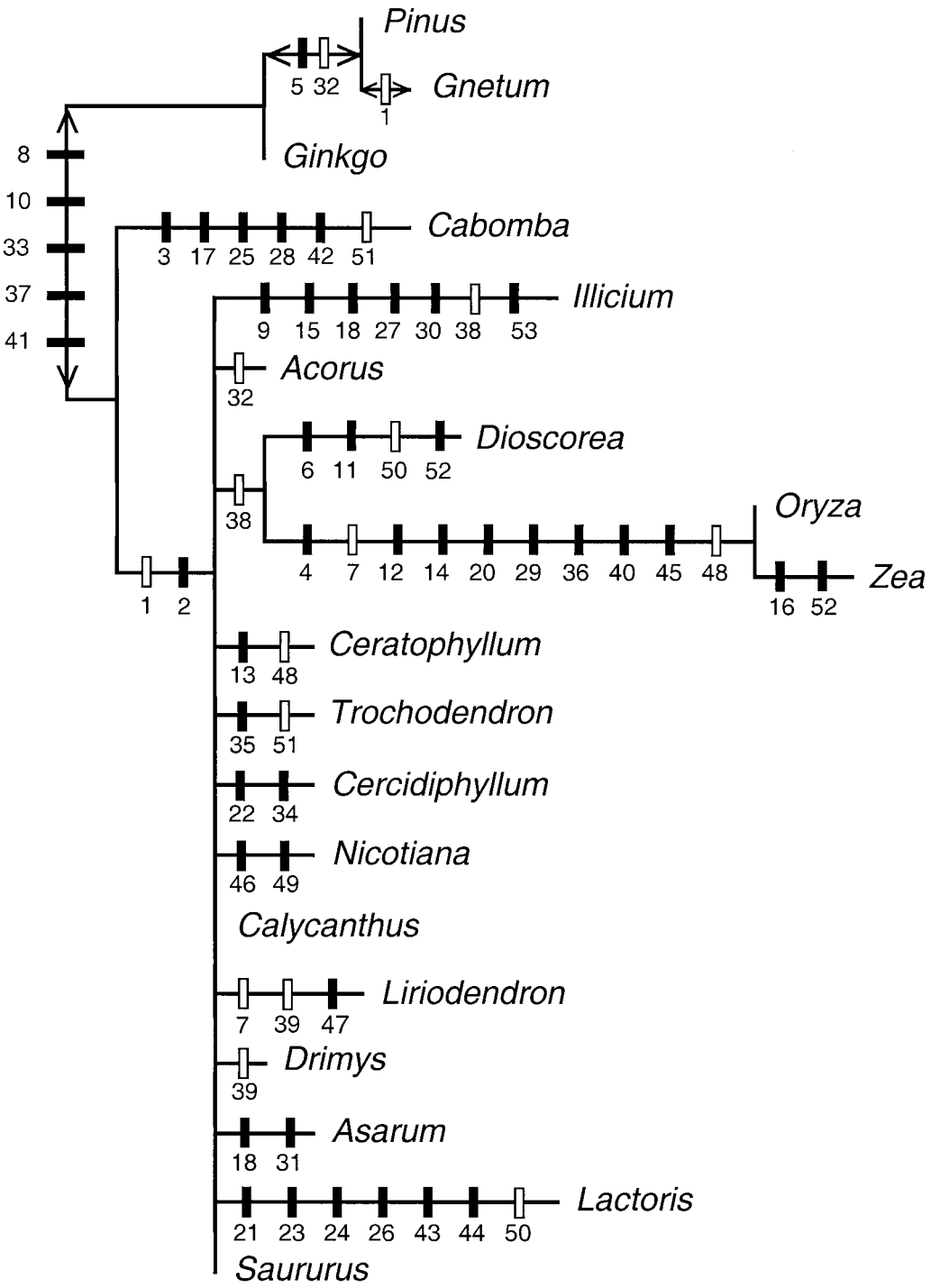


Fig. 7. Indel characters mapped onto the maximum parsimony tree shown in Fig. 6 (reduced to show only branches supported by indels) using ACCTRAN optimization. Numbered bars on each branch indicate the indel events inferred along them (the numbered labels correspond to those in Table 3). Indels observed to be homoplasious on the tree in Fig. 6. are represented by hollow bars. For outgroup taxa, brackets around bars indicate that they reflect only nonambiguous or nonunique indel events (see text).

rejected as the sister-group of the rest of the living angiosperms in favor of *Cabomba* (Fig. 9a). However, when *Amborella* and *Nymphaea odorata* were added together, *Amborella* was moderately well supported as the sister-group of all other angiosperms, and the two major water lily lineages, represented here by *Cabomba* and *Nymphaea*, were together inferred to be the sister-group of the remaining angiosperms (Fig. 9b). In all analyses *Illicium* was depicted as the next most basal angiosperm lineage, after *Amborella* and the water lilies. The two analyses that included *Amborella* also disagreed over relationships inferred among the major seed plant groups (Fig. 9). However, in both cases

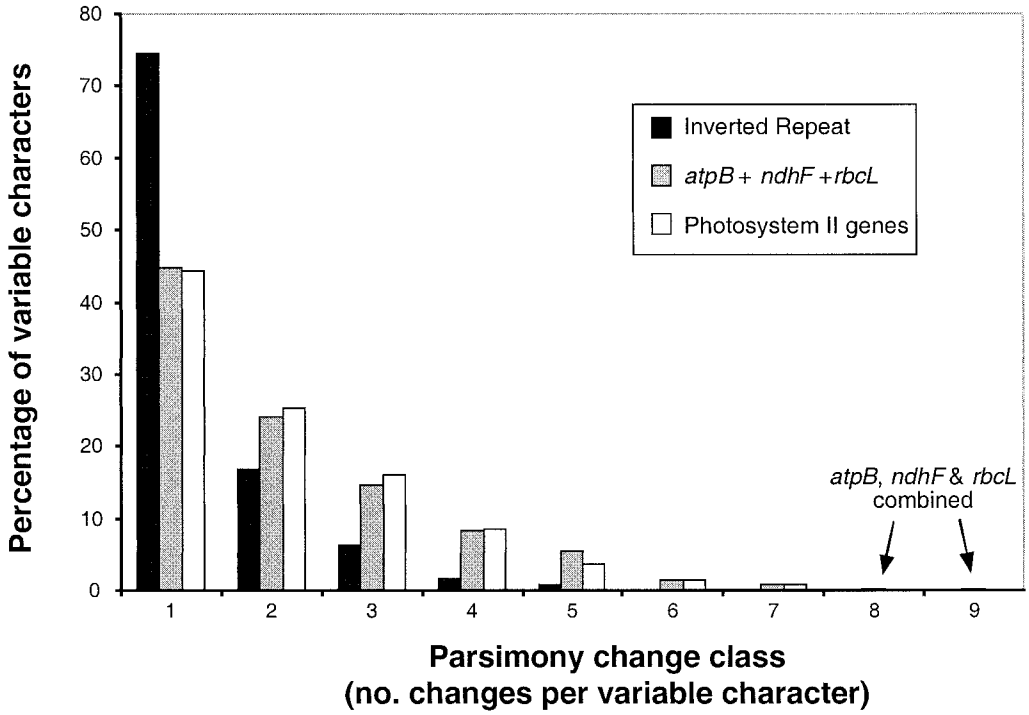


Fig. 8. Distribution of parsimony-change classes for three partitions of the data. The changes were those inferred on the angiosperm subtree in Fig. 6. Only nucleotide characters were considered. These are minimal change estimates: denser taxon sampling should detect more multiple changes. Note that nucleotide sites sometimes are multistate characters, so sites with two or three changes will not all be homoplastic.

analyses that excluded outgroup taxa yielded the same underlying ingroup topology as that shown in Figs. 6 and 9 (results not shown).

DISCUSSION

Utility of the new primers—Sampling error on short branches is an important source of ambiguity in phylogenetic reconstruction (e.g., Rodrigo et al., 1993; Page, 1996; Graham et al., 1998). The large number of characters in our study, about a tenth of the entire chloroplast genome (nearly 13.4 kb of unaligned coding and intron sequence in *Nicotiana*), provides substantial additional leverage for overcoming sampling error. The number of characters that can be examined has been somewhat limited by the availability of suitable PCR and sequencing primers, and the difficulty of obtaining large amounts of sequence. Our major goals were to present new sets of primers for plant molecular systematics, and to describe their application and utility in studying deep phylogenetic divisions in the flowering plants. Ten of 13 internal branches in our initial sampling of basal angiosperm lineages have $>$ 70% support from bootstrap analysis (Fig. 6). The primers described here thus constitute a valuable new set of tools for inference of basal angiosperms (and other land plant) relationships. Nonetheless, it should be emphasized that high bootstrap values may sometimes reflect erroneous relationships in areas of inconsistency, rather than the accuracy of particular relationships (see below).

Quality of the data—Characters with a conservative evolutionary rate are expected be more resistant to long-branch attraction (Felsenstein, 1983). The regions examined here are at least as slowly evolving as those previously being used to assess angiosperm phylogeny, and in several cases substantially slower (Table 5; Fig. 8). The retention index, RI, has been used as a criterion for assessing the relative informativeness of each character, or as a measure of phylogenetic signal (Farris, 1989; Savolainen et al., in press). The Photosystem II data have very similar RI values to the combined *atpB* + *ndhF* + *rbcL* data (Table 5), and so by this measure the *psb* genes are comparable to the other single-copy genes examined. The IR data have by far the highest RI values (Table 5), and so individual IR characters would be expected to be on average more reliable than those from any of the other classes of data considered here.

A large gamma shape parameter indicates that all sites evolve at essentially the same rate (Swofford et al., 1996). Very small alpha values, such as we find here (Table 5), can indicate highly asymmetrical distributions of rates, with most sites changing very little or not at all (see Swofford et al., 1996) and a few sites changing more frequently. We suggest, therefore, that with slowly evolving sequences, very low alpha values can be thought of as roughly approximating an equal-rates model, at least within a maximum parsimony framework. This is in line with the findings of Felsenstein (1981, 1983) that when most characters change sufficiently slowly they may be equally weighted, even though they do not all actually change with equal probability, provided that the overall rate of change is very low (M. Sanderson, personal communication).

Figure 8 indicates that the IR data set most closely approximates an equal-rates model under parsimony, since most of its variable characters fall in only one parsimony-change class, the class with only one change inferred per character. This also largely accounts for the very low amount of homoplasy in-

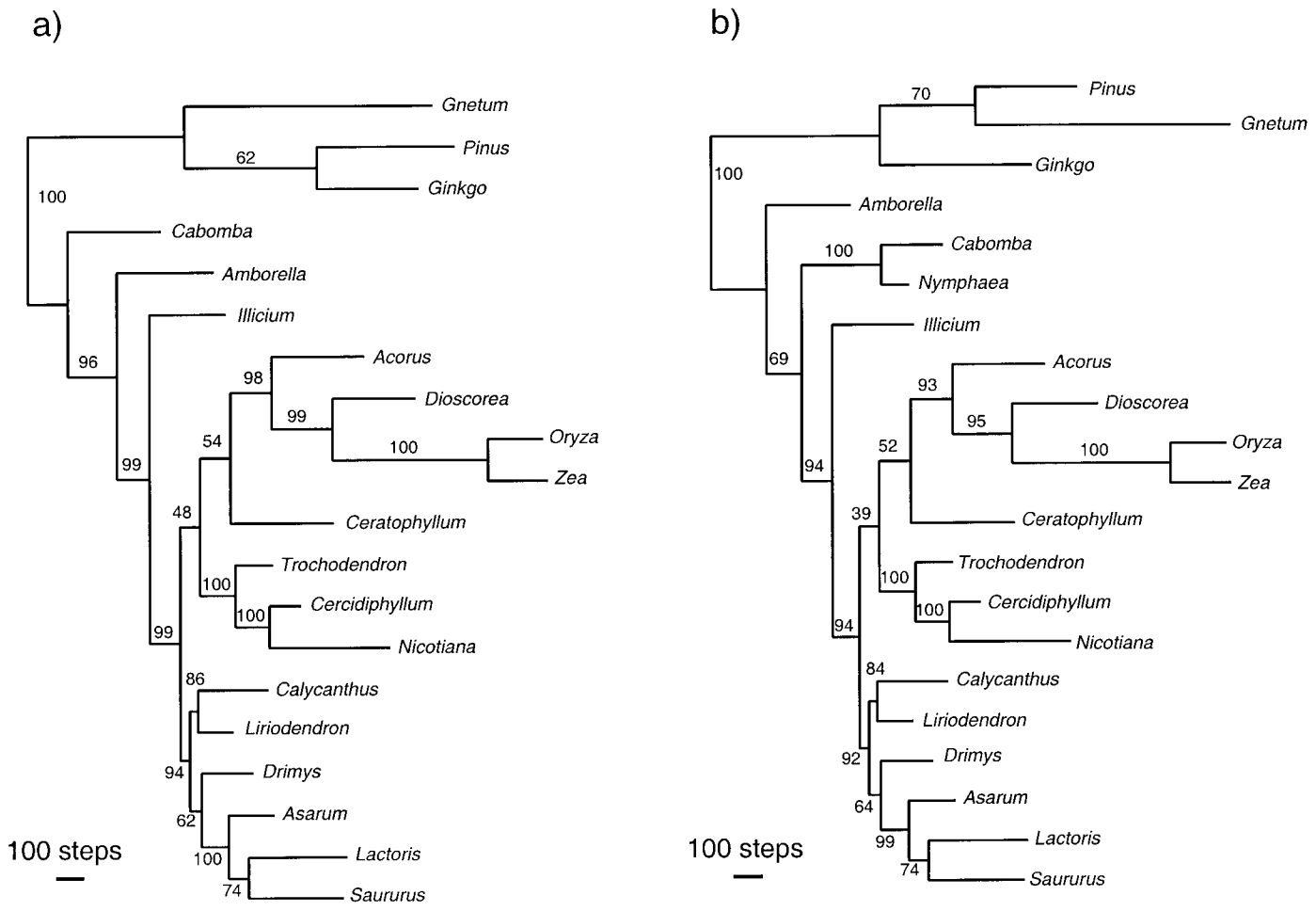


Fig. 9. Single most parsimonious trees found using combined sequence, intron, and indel data from 17 chloroplast genes, with one or two basal angiosperms added to the preliminary data set. (a) Only *Amborella trichopoda* added (tree length = 9105 steps, CI = 0.601, RI = 0.460). (b) Both *Amborella* and *Nymphaea odorata* added (tree length = 9261 steps, CI = 0.595, RI = 0.478). In both cases the same underlying ingroup topology was seen when outgroups were excluded (results not shown). Branch lengths are proportional to the total amount of inferred change (computed using ACCTRAN optimization). Scale bars are included. Bootstrap support is noted beside each branch.

ferred for this class of data. Indeed, <3% of variable IR characters are inferred to change three times or more across the entire tree. By comparison, 17 and 14% of variable characters change more than three times for the combined *atpB* + *ndhF* + *rbcL* data and the combined *psb* data, respectively (Fig. 8). The total amount of change per site is likely to be an underestimate, because of undetected homoplasy, but this effect is thought to be minor for sites that have experienced few changes (Wakely, 1993), as is the case for most of the variable chloroplast characters we examined.

Long-branch attraction and basal angiosperm relationships—Attraction between exceptionally long branches neighboring short internodes can become stronger as more characters (e.g., Felsenstein, 1978; Huelsenbeck, 1995) or taxa (Swofford and Poe, 1999) are examined. This phenomenon should thus be more apparent with molecular data sets than with morphological studies, by dint of their greater size. The observed strong conflicts among different subsets of our data, or at different levels of taxonomic sampling, may also be a consequence of rather severe long branches. Our suggestion that strongly divergent results can also be a hallmark of long-branch attraction is not necessarily in conflict with the standard view that it results in strong convergence to a single wrong answer: long-branch attraction is poorly understood for more than about four or five taxa (e.g., Kim, 1996).

Several candidate long-branch effects are apparent in our study using this criterion. One concerns outgroup relationships. In our initial taxon sampling, the two single-copy data partitions converged strongly on different arrangements of the four major seed-plant groups considered, with the angiosperms grouping strongly with either *Ginkgo* or *Gnetum.* They also clashed over the position of *Lactoris* within Piperales. In each case the conflicting relationship was supported by > 70% of bootstrap replicates (see Results). The long branch associated with *Lactoris* is implied not only by the number of substitutions inferred on the maximum parsimony tree (Fig. 6; note that *Saururus* is almost as long), but also by the large number of indel events inferred along this terminal branch (Fig. 7). The significant result with the test of Farris et al. (1994) (with *Lactoris* and the outgroups included; see Results) may also reflect substantial ''saturation'' or noise on these long branches (see Graham et al., 1998). When the outgroups were ignored, all of the analyses involving the 17 combined genes found the

same underlying relationships among the angiosperms (see Results). However, when outgroups were included in the analyses, an additional candidate long-branch problem involved the inferred root of the angiosperms.

The root of the angiosperms—For the initial set of analyses involving 16 core angiosperm taxa and three outgroups, we found strong support for the first two basal splits, represented by *Cabomba* and *Illicium*, respectively. *Cabomba* was used to represent the water lilies and *Illicium* represents a woody magnoliid group distinct from the core woody magnoliids (Magnoliales, Laurales, and Winteraceae). The root split at the water lilies was further supported by two indels here (Table 3; Fig. 7), and by a single indel in the very slowly evolving chloroplast ITS region (Goremykin et al., 1996).

The determination of the root node of a large taxon, such as the angiosperms, is always conditional on increased taxon sampling (Sanderson, 1996). Since our submission of this paper, a fast-paced series of developments in basal angiosperm phylogeny has taken place, in studies that employ a variety of genes and levels of taxon sampling (Mathews and Donoghue, 1999; Qiu et al., 1999; Soltis, Soltis, and Chase, 1999; Parkinson, Adams, and Palmer, 1999; Graham et al., in press; S. W. Graham and R. G. Olmstead, unpublished data). All of these studies, and the current study, support the idea that the water lilies, and next, *Illicium* and relatives, represent successively emerging basal lineages close to the base of the flowering plants. However, perhaps the most significant new discovery has been that the New Caledonian species *Amborella trichopoda* (Amborellaceae) may constitute the sister-group of the rest of the angiosperms. These exciting results have been widely commented upon in the popular media and have led many botanists to view the problem of rooting the angiosperms to be essentially solved.

We therefore decided to anticipate a future, more detailed study on the root of the angiosperms and add *Amborella* and one additional major water lily lineage (*Nymphaea odorata*) to our preliminary taxon sampling. Using only slightly different levels of taxon sampling (*Nymphaea* included or excluded), *Cabomba* or *Amborella* were each strongly supported as candidate sister-groups of the rest of the angiosperms (Fig. 9). The rooting at *Cabomba* (Fig. 9A), inferred when only *Amborella* was added, also conflicts strongly with bootstrap analyses reported in Mathews and Donoghue (1999), Qiu et al. (1999), Soltis, Soltis, and Chase (1999) and Parkinson, Adams, and Palmer (1999). Additional conflict was seen in outgroup relationships in the analyses involving this additional taxon sampling (Fig. 9).

Our analyses thus suggest that it is premature to place confidence in the *Amborella* rooting of the angiosperms, in this or other published studies with fewer characters available for analysis. The extant seed-plant groups are separated by very long branches that cannot be broken apart by the inclusion of additional intermediate taxa, because these are now extinct. A number of basal angiosperm lineages have similarly long branches. This fact alone should serve to give pause to the idea that the rooting of the angiosperms has been solved (see also Niklas, Crepet, and Nixon, 1999).

Studies in progress will attempt to address whether the result here is an expression of deeper problems with the widely reported *Amborella* rooting. Of the two lineages competing for position at the base of the angiosperms in our analyses, *Amborellaceae* is a monotypic family, and *Nymphaea* and *Cabomba* represent each of the two major lineages of water lilies (Les et al., 1999). Therefore, it is unlikely that the branches leading to these basal angiosperm lineages will be broken up more than we have done in this preliminary study (Fig. 9), even with substantial additional taxon sampling. The conflicting results described here concerning the rooting of the angiosperms thus may not be settled by additional taxon sampling alone.

Conclusion—In our initial taxon sampling the placement of the root of the angiosperms between water lilies and the other exemplar angiosperms was found by all three major data partitions we examined, and the combined data and two of the three data partitions supported this strongly. Most of the remaining clades were also well supported. The genes we used were carefully chosen to survey a large number of slowly evolving characters, using new primers that worked well across a broad range of seed-plant taxa. The loci examined have low synonymous substitution rates, low homoplasy, and approximate an equal-rates model under parsimony. In combination with other chloroplast data they provide about an order of magnitude more high-quality characters than the landmark *rbcL* study of Chase et al. (1993). We were also able to demonstrate several candidate cases where long-branch attraction may contribute to erroneous phylogenetic inference, including inference of the root node of the angiosperms: with slightly different taxon samplings two different root nodes were found for the angiosperms, one in strong conflict with published rootings.

LITERATURE CITED

ANGIOSPERM PHYLOGENY GROUP (APG). 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85: 531–553.

BOCK, R., R. HAGEMANN, H. KÖSSEL, AND J. KUDLA. 1993. Tissue- and stage-specific modulation of RNA editing of the *psbF* and *psbL* transcript from spinach plastids—a new regulatory mechanism? *Molecular and General Genetics* 240: 238–244.

BOWE, L. M., AND C. W. DEPAMPHILIS. 1996. Effects of RNA editing and gene processing on phylogenetic reconstruction. *Molecular Biology and Evolution* 13: 1159–1166.

CHASE, M. W., D. E. SOLTIS, R. G. OLMSTEAD, ET AL. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80: 528–580.

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. *Molecular Biology and Evolution* 14: 56–68.

CRANE, P. R. 1993. Time for the angiosperms. *Nature* 366: 631–632.

———, E. M. FRIIS, AND K. R. PEDERSEN. 1995. The origin and early diversification of angiosperms. *Nature* 374: 27–33.

CREPET, W. L. 1998. The abominable mystery. *Science* 282: 1653–1654.

DONOGHUE, M. J., AND J. A. DOYLE. 1989. Phylogenetic studies of seed plants and angiosperms based on morphological characters. *In* B. Fernholm, K. Bremer, and H. Jörnvall [eds.], The hierarchy of life, 181–193. Elsevier, Amsterdam, The Netherlands.

———, AND M. J. SANDERSON. 1992. The suitability of molecular and morphological evidence in reconstructing plant phylogeny. *In* P. S. Soltis, D. E. Soltis, and J. J. Doyle [eds.], Molecular systematics of plants, 340–368. Chapman and Hall, New York, New York, USA.

DOWNIE, S. R., R. G. OLMSTEAD, G. ZURAWSKI, D. E. SOLTIS, P. E. SOLTIS, J. C. WATSON, AND J. D. PALMER. 1991. Six independent losses of the chloroplast *rpl*2 intron in dicotyledons: molecular and phylogenetic implications. *Evolution* 45: 1245–1259.

DOYLE, J. A. 1998. Phylogeny of the vascular plants. *Annual Review of Ecology and Systematics* 29: 567–599.

———, AND M. J. DONOGHUE. 1993. Phylogenies and angiosperm diversification. *Paleobiology* 19: 141–167.

ENDRESS P. K., AND A. IGERSHEIM, 1997. Gynoecium diversity and systematics of the Laurales. *Botanical Journal of the Linnean Society* 125: 93–168.

ENGELS, W. 1993. Amplify (version 1.2). Computer program and documentation. Genetics Department, University of Wisconsin, Madison, Wisconsin, USA.

FARRIS, J. S. 1989. The retention index and the rescaled consistency index. *Cladistics* 5: 417–419.

———, M. KÄLLERSJÖ, A. G. KLUGE, AND C. BULT. 1994. Testing significance of incongruence. *Cladistics* 10: 315–319.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* 27: 401–410.

———. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society* 16: 183–196.

———. 1983. Parsimony in systematics: biological and statistical issues. *Annual Review of Ecology and Systematics* 14: 313–333.

———. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783–791.

———. 1995. PHYLIP (Phylogeny Inference Package) version 3.5c. Computer programs and documentation. Department of Genetics, University of Washington, Seattle, Washington, USA.

———, AND H. KISHINO. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. *Systematic Biology* 42: 193–200.

FROHLICH, M. W. 1999. MADS about Gnetales. *Proceedings of the National Academy of Sciences, USA* 96:8811–8813.

FREYER, R., M.-C. KIEFER-MEYER, AND H. KÖSSEL. 1997. Occurrence of plastid RNA editing in all major lineages of land plants. *Proceedings of the National Academy of Sciences, USA* 94: 6285–6290.

GOREMYKIN, V., V. BOBROVA, J. PAHNKE, A. TROITSKY, A. ANTONOV, AND W. MARTIN. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbcL* data do not support gnetalean affinities of angiosperms. *Molecular Biology and Evolution* 13: 383–396.

GRAHAM, S. W., J. R. KOHN, B. R. MORTON, J. E. ECKENWALDER, AND S. C. H. BARRETT. 1998. Phylogenetic congruence and discordance among one morphological and three molecular data sets from Pontederiaceae. *Systematic Biology* 47: 545–567.

———, AND R. G. OLMSTEAD. 2000. Evolutionary significance of an unusual chloroplast DNA inversion found in two basal angiosperm lineages. *Current Genetics* 37: 183–188.

———, P. A. REEVES, A. C. E. BURNS, AND R. G. OLMSTEAD. In press. Microstructural changes in noncoding chloroplast DNA: interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. *International Journal of Plant Sciences.*

GRUISSEM, W., AND J. C. TONKYN. 1993. Control mechanisms of plastid gene expression. *Critical Reviews in Plant Sciences.* 12: 19–55.

HALEY, J., AND L. BOGORAD. 1990. Alternative promoters are used for genes within maize chloroplast polycistronic units. *Plant Cell* 2: 323–333.

HENDY, M. D., AND D. PENNY. 1989. A framework for the quantitative study of evolutionary trees. *Systematic Zoology* 38: 297–309.

HICKEY, L. J., AND D. W. TAYLOR. 1996. Origin of the angiosperm flower. *In* D. W. Taylor and L. J. Hickey [eds.], Flowering plant origin, evolution, and phylogeny, 176–231. Chapman and Hall, New York, New York, USA.

HILLIS, D. M., AND J. J. BULL. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology* 42: 182–192.

HOOT, S. B., A. CULHAM, AND P. R. CRANE. 1995. The utility of *atpB* gene sequences in resolving phylogenetic relationships: comparison with *rbcL* and 18S ribosomal DNA sequences in the Lardizabalaceae. *Annals of the Missouri Botanical Garden* 82: 194–207.

HUELSENBECK, J. P. 1995. Performance of phylogenetic methods in simulation. *Systematic Biology* 44: 17–48.

IGERSHEIM, A., AND P. K. ENDRESS. 1997. Gynoecium diversity and systematics of the Magnoliales and winteroids. *Botanical Journal of the Linnean Society* 124: 213–271.

———, AND ———. 1998. Gynoecium diversity and systematics of the paleoherbs. *Botanical Journal of the Linnean Society* 127: 289–370.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. *Systematic Biology* 45: 363–374.

KIM, K.-J., AND R. K. JANSEN. 1995. *ndhF* sequence evolution and the major clades in the sunflower family. *Proceedings of the National Academy of Sciences, USA* 92: 10379–10383.

KUDLA, J., G. L. IGLOI, M. METZLAFF, R. HAGEMANN, AND H. KÖSSEL. 1992. RNA editing in tobacco chloroplasts leads to the formation of a translatable *psbL* mRNA by a C to U substitution within the initiation codon. *EMBO Journal* 11: 1099–1103.

LES, D. H., E. L. SCHNEIDER, D. J. PADGETT, P. E. SOLTIS, D. E. SOLTIS, M. ZANIS. 1999. Phylogeny, classification and floral evolution of water lilies (Nymphaeaceae, Nymphaeales): a synthesis of non-molecular, *rbcL*, *matK*, and 18S rDNA data. *Systematic Botany* 24: 28–46.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Massachusetts, USA.

LOCONTE, H. 1996. Comparison of alternative hypotheses for the origin of the angiosperms. *In* D. W. Taylor and L. J. Hickey [eds.], Flowering plant origin, evolution, and phylogeny, 267–285. Chapman and Hall, New York, New York, USA.

———, AND D. W. STEVENSON. 1991. Cladistics of the Magnoliidae. *Cladistics* 7: 267–296.

MAIER, R. M., K. NECKERMANN, G. L. IGLOI, AND H. KÖSSEL. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing. *Journal of Molecular Biology* 251: 614–628.

MATHEWS, S., AND M. J. DONOGHUE. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286: 947–950.

NANDI, O., M. W. CHASE, AND P. K. ENDRESS. 1998. A combined cladistic analysis of angiosperms using *rbcL* and non-molecular data sets. *Annals of the Missouri Botanical Garden* 85: 137–212.

NIKLAS, K. J., W. L. CREPET, AND K. C. NIXON. 1999. Early plant history: something borrowed, something new? *Science* 285:1673.

OLMSTEAD, R. G., AND J. D. PALMER. 1994. Chloroplast DNA systematics: a review of methods and data analysis. *American Journal of Botany* 81: 1205–1224.

———, AND P. A. REEVES. 1995. Evidence for the polyphyly of the Scrophulariaceae based on chloroplast *rbcL* and *ndhF* sequences. *Annals of the Missouri Botanical Garden* 82: 176–193.

———, AND J. A. SWEERE. 1994. Combining data in phylogenetic systematics: an empirical approach using three molecular data sets in the Solanaceae. *Systematic Biology* 43: 467–481.

PAGE, R. D. M. 1996. On consensus, confidence, and ''total evidence.'' *Cladistics* 12: 83–92.

PARKINSON, C. L., K. L. ADAMS, AND J. D. PALMER. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. *Current Biology* 9:1485–1488.

QIU, Y.-L., J. LEE, F. BERNASCONI-QUADRONI, D. E. SOLTIS, P. E. SOLTIS, M. ZANIS, E. A. ZIMMER, Z. CHEN, V. SAVOLAINEN, AND M. W. CHASE. 1999. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. *Nature* 402: 404–407.

RAMBAUT, A. 1998. Se-Al (Sequence Alignment Editor Version 1.0 alpha 1). Computer program and documentation. Department of Zoology, University of Oxford, UK.

RODRIGO, A. G., M. KELLY-BORGES, P. R. BERGQUIST, AND P. L. BERGQUIST. 1993. A randomisation test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. *New Zealand Journal of Botany* 31: 257–268.

SANDERSON, M. J. 1996. How many taxa must be sampled to identify the root node of a large clade? *Systematic Biology* 45: 168–173.

SAVOLAINEN, V., M. W. CHASE, S. B. HOOT, C. M. MORTON, D. E. SOLTIS, C. BAYER, M. FAY, A. Y. DE BRUIJN, S. SULLIVAN, AND Y.-L. QIU. 2000. Phylogenetics of flowering plants based on a combined analysis of plastid *atpB* and *rbcL* gene sequences. *Systematic Biology* 49: 306–362.

SOLTIS, P. E., D. E. SOLTIS, AND M. W. CHASE. 1999. Angiosperm phylogeny inferred from multiple chloroplast genes as a tool for comparative biology. *Nature* 402: 402–404.

SWOFFORD, D. L. 1999. PAUP* 4.0 (beta version). Computer program and documentation. Sinauer, Sunderland, Massachusetts, USA.

———, G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. *In* D. M. Hillis, C. Moritz, and B. K. Mable [eds.], Molecular systematics, 2nd ed., 407–514. Sinauer, Sunderland, Massachusetts, USA.

———, AND S. POE. 1999. Taxon sampling revisited. *Nature* 398:299–300.

SYTSMA, K. J., AND D. A. BAUM. 1996. Molecular phylogenies and the diversification of the angiosperms. *In* D. W. Taylor and L. J. Hickey

[eds.], Flowering plant origin, evolution, and phylogeny, 314–340. Chapman and Hall, New York, New York, USA.

TAYLOR, D. W., AND L. J. HICKEY. 1992. Phylogenetic evidence for the herbaceous origin of angiosperms. *Plant Systematics and Evolution* 180: 137–156.

THOMPSON, J. D., D. G. HIGGINS, AND T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673–4680.

TUCKER, S. C., AND A. W. DOUGLAS. 1996. Floral structure, development, and relationships of paleoherbs: *Saruma, Cabomba, Lactoris*, and selected Piperales. *In* D. W. Taylor and L. J. Hickey [eds.], Flowering plant origin, evolution, and phylogeny, 141–175. Chapman and Hall, New York, New York, USA.

WAKELY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *Journal of Molecular Evolution* 37: 613–623.

WOLFE, K. H., W.-H. LI, AND P. M. SHARP. 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences, USA* 84: 9054–9058.

YAO, W. B., B. Y. MENG, M. TANAKA, AND M. SUGIURA. 1989. An additional promoter within the protein-coding region of the *psbD-psbC* gene cluster in tobacco chloroplast DNA. *Nucleic Acids Research* 17: 9583–9591.

ZAITA, N., K. TORAZAWA, K. SHINOZAKI, AND M. SAGIURA. 1987. *Trans* splicing in vivo: joining of transcripts from the 'divided' gene for ribosomal protein S12 in the chloroplast of tobacco. *FEBS Letters* 210: 153–156.

ZURAWSKI, G., M. T. CLEGG, AND A. H. D. BROWN. 1984. The nature of nucleotide sequence divergence between barley and maize chloroplast DNA. *Genetics* 106: 735–749.

APPENDIX. Primer statistics.

Region:	3′ *rps12*, *rps7*, *ndhB* and *trnL*																
Gene:	3′ *rps12*				*rps7*	*ndhB*											*trnL*
Fragment:	exon 1		intron			exon 1						intron		exon 2			
Primer	lF	2F[a]	3R	4F[a]	5R	6F/6R	7F/7R	8R	9F	B10R	10F/10R	C11R	11F/11R	12F	13F	14R	15R
Product used for sequencing[b]	A	A	A	A	A	B/A	B/A	B	C	B	C/*	C	D/C	(D)	D	D	*
Position of 5′-most base[c]																	
(1) In gene (or exon/intron)	10	212	349	523	100	422	85	320	587	712	751	295	475	25	170	682	21
(2) In entire sequence	10	212	581	755	947	1269	1678	1913	2180	2305	2344	2665	2845	3074	3219	3731	4365

[a] Straddles intron/exon boundary.
[b] PCR product used to sequence with this primer: A = lF/7R; B = 6F/10R; C = 9F/13R; D = 11F/14R; parentheses = used as an alternate sequencing primer; * = not used to sequence.
[c] Reference sequence = *Nicotiana tabacum*.

Gene/region:	*rpl2*							
Fragment:	exon 1			intron			exon 2	
Primer	20F	B20F	21F[a]	22R	23F	24F/24R[a]	B25R	25R
Product used for sequencing[b]	E	(E)	E	E	E	E	(E)	E
Position of 5′-most base[c]								
(1) In exon/intron fragment	115	229	380	184	421	664	262	352
(2) In entire sequence	115	229	380	575	812	1055	1318	1408

[a] Straddles intron/exon boundary.
[b] PCR product used to sequence with this primer: E = 20F/25R; parentheses = used as an alternate sequencing primer.
[c] Reference sequence = *Nicotiana tabacum*.

Region:	*psbD/psbC*												
Gene:	*psbD*					*psbC*							
Primer	40F	41R	42F	44F/B44R	45R[a]	46F	47R	B47F/B47R	48F	49R	50F	51R	B51R
Product used for sequencing[b]	F	F	F	F/F	F	G	*	(G)/G	G	G	G	G	(G)
Position of 5′-most base[c]													
(1) In gene	59	301	368	775	1010	94	424	452	481	856	935	1256	1294
(2) In entire sequence	59	301	368	775	1010	1103	1433	1461	1490	1865	1944	2265	2303

[a] Primer 45R also starts at first base pair of *psbC*.
[b] PCR product used to sequence with this primer: F = 40F/47R; G = 44F/B51R; parentheses = used as an alternate sequencing primer; * = not used to sequence.
[c] Reference sequence = *Nicotiana tabacum*.

Region:	*psbE/psbF/psbL/psbJ*						
Gene:	*psbE*		*psbF*		*psbL*	*psbJ*	
Primer	55F	B55F	B56F	56R	B57F	B58R	58R
Product used for sequencing[a]	H	(H)	H	H	(H)	(H)	H
Position of 5′-most base[b]							
(1) In gene	1	61	85	89	68	8	92
(2) In entire sequence	1	61	346	350	471	652	736

[a] PCR product used to sequence with this primer: H = 55F/5R; parentheses = used as an alternate sequencing primer.
[b] Reference sequence = *Nicotiana tabacum*.

Region:	*psbB*, *psbT*, *psbN*, and *psbH*															
Gene:	*psbB*											*psbT*		*psbN*	*psbH*	
Primer	60F	B60F	B61F	61R	63F	B64R	65F	B66R	66R	67F	*B68R	68R	69F	70R	B71R	71R
Product used for sequencing[a]	I	I	I	I	I	I	J	I	*	J	(J)	J	J	J	(J)	J
Position of 5′-most base[b]																
(1) In gene	1	76	263	332	682	824	1076	1321	1337	1427	1485	1	56	32[c]	85	110
(2) In entire sequence	1	76	263	332	682	824	1076	1321	1337	1427	1485	1728	1783	1979	2233	2258

[a] PCR product used to sequence with this primer: I = 60F/66R; J = 65F/71R; parentheses = used as an alternate sequencing primer.
[b] Reference sequence = *Nicotiana tabacum*.
[c] From 5′ end of gene (Note: *psbN* is on the opposite strand to the *psbB*/*psbT*/*psbH* operon).