Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns

Robert K. Jansen[†‡], Zhengqiu Cai[†], Linda A. Raubeson[§], Henry Daniell[¶], Claude W. dePamphilis[‖], James Leebens-Mack[††], Kai F. Müller[‖‡‡], Mary Guisinger-Bellian[†], Rosemarie C. Haberle[†], Anne K. Hansen[†], Timothy W. Chumley[†], Seung-Bum Lee[¶], Rhiannon Peery[§], Joel R. McNeal[††], Jennifer V. Kuehl[§§], and Jeffrey L. Boore[§§¶¶]

[†]Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas, Austin, TX 78712; [§]Department of Biological Sciences, Central Washington University, Ellensburg, WA 98926; [¶]Department of Molecular Biology and Microbiology, Biomolecular Science, University of Central Florida, Orlando, FL 32816; [‖]Department of Biology and Institute of Molecular and Evolutionary Genetics, Huck Institutes of Life Sciences, Pennsylvania State University, University Park, PA 16802; [††]Department of Plant Biology, University of Georgia, Athens, GA 30602; [§§]Department of Energy Joint Genome Institute and Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598; and [‡‡]Nees Institute for Biodiversity of Plants, University of Bonn, 53115 Bonn, Germany

Communicated by Peter H. Raven, Missouri Botanical Garden, St. Louis, MO, September 25, 2007 (received for review July 6, 2007)

Angiosperms are the largest and most successful clade of land plants with >250,000 species distributed in nearly every terrestrial habitat. Many phylogenetic studies have been based on DNA sequences of one to several genes, but, despite decades of intensive efforts, relationships among early diverging lineages and several of the major clades remain either incompletely resolved or weakly supported. We performed phylogenetic analyses of 81 plastid genes in 64 sequenced genomes, including 13 new genomes, to estimate relationships among the major angiosperm clades, and the resulting trees are used to examine the evolution of gene and intron content. Phylogenetic trees from multiple methods, including model-based approaches, provide strong support for the position of *Amborella* as the earliest diverging lineage of flowering plants, followed by Nymphaeales and Austrobaileyales. The plastid genome trees also provide strong support for a sister relationship between eudicots and monocots, and this group is sister to a clade that includes Chloranthales and magnoliids. Resolution of relationships among the major clades of angiosperms provides the necessary framework for addressing numerous evolutionary questions regarding the rapid diversification of angiosperms. Gene and intron content are highly conserved among the early diverging angiosperms and basal eudicots, but 62 independent gene and intron losses are limited to the more derived monocot and eudicot clades. Moreover, a lineage-specific correlation was detected between rates of nucleotide substitutions, indels, and genomic rearrangements.

angiosperm evolution | molecular evolution

Angiosperms, the largest clade of land plants with >250,000 species, experienced rapid radiation soon after their first appearance in the fossil record (1). As a result, flowering plants exhibit incredible diversity in habit, morphology, anatomy, physiology, and reproductive biology. This variation has presented major challenges to biologists interested in the origin and evolution of these traits, and resolving these issues critically depends on having a well resolved and strongly supported phylogenetic framework. Over the past 20 years, numerous phylogenetic studies have used both morphological and molecular data to assess relationships among the major clades (reviewed in ref. 2), resulting in a widely accepted classification of angiosperms with 45 orders and 457 families (3).

For nearly two decades, most phylogenetic analyses of angiosperms have relied on DNA sequences of one to several genes from the plastid, mitochondrial, and nuclear genomes (reviewed in ref. 2). Despite these intensive efforts there are still uncertainties regarding relationships among several major clades throughout angiosperms, including the earliest diverging lineages. Recent studies support the placement of *Amborella* sister to all remaining angiosperms, but support is often low. *Amborella* has also been placed with waterlilies (Nymphaeales) in a clade sister to other angiosperms (4–7). In many studies, resolution of relationships among *Amborella*, Nymphaeales, and the rest of angiosperms varies with both the phylogenetic method and gene and taxon sampling (e.g., refs. 4 and 6–8). For example, using eight markers from all three genomes (6), plastid data support *Amborella* as the sole sister group of the remaining angiosperms, whereas mitochondrial genes support *Amborella* plus Nymphaeales as sister to other angiosperms. A more recent analysis of 17 plastid genes and the nuclear gene phytochrome C (*PHYC*) found weak support for *Amborella* as the basal-most angiosperm lineage followed by a strongly supported clade including Nymphaeales and Hydatellaceae (9). Most phylogenetic trees (e.g., refs. 7, 10–11) show Austrobaileyales to be the next-diverging lineage, whereas relationships among Ceratophyllaceae, magnoliids, Chloranthales, monocots, and eudicots are typically weakly supported and fluctuate depending on the markers and phylogenetic methods used.

Several recent phylogenetic studies have used 61 protein-coding genes from completely sequenced plastid genomes (5, 12–16). These genome-scale analyses have the potential to provide the data necessary to resolve relationships among the major clades of angiosperms; however, genome-scale phylogenetic studies can be susceptible to long-branch artifacts (17, 18) when taxon sampling is sparse (5, 19–22). To resolve relationships among the major clades of angiosperms, we present phylogenetic analyses based on a greatly expanded number of genes and taxa, using 81 genes from 64 sequenced plastid genomes, by far the most extensive data matrix applied to this issue. In addition to providing a fully resolved and strongly supported phylogenetic tree of the major clades, we use this tree to examine the evolution of gene and intron content and correlations between rates of insertions/deletions, nucleotide substitutions, and genomic rearrangements in plastid genomes.

Author contributions: R.K.J. designed research; Z.C., L.A.R., H.D., C.W.d., J.L.-M., M.G.-B., R.C.H., A.K.H., T.W.C., S.-B.L., R.P., J.R.M., J.V.K., and J.L.B. performed research; R.K.J., Z.C., L.A.R., J.L.-M., and K.F.M. analyzed data; and R.K.J. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession numbers are provided in SI Table 1). The aligned data matrix is available at http://chloroplast.cbio.psu.edu.

[‡]To whom correspondence should be addressed. E-mail: jansen@mail.utexas.edu.

[¶¶]Present addresses: Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547; and University of California, 3060 VLSB, Berkeley, CA 94702.

This article contains supporting information online at www.pnas.org/cgi/content/full/0709121104/DC1.

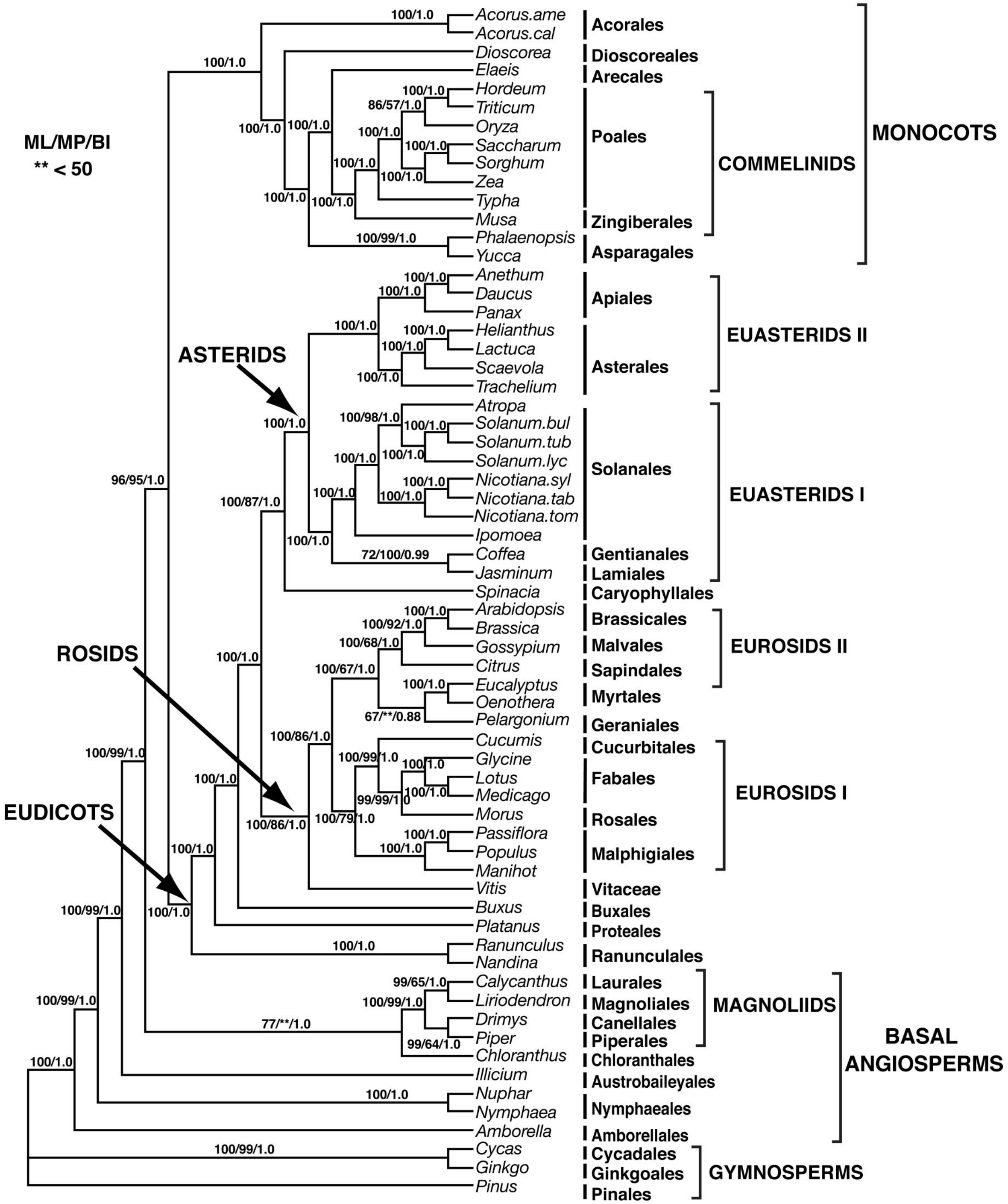


Fig. 1. ML tree of 64 taxa based on 81 plastid gene sequences. The tree has a −lnL of 886368.804.118. Support values for ML, MP, and BI are provided at the nodes. Where ML and MP numbers >50% are identical, only one number is provided. Names for major clades follow angiosperm phylogeny group II classification (3).

Results

Phylogenetic analyses were performed on a 64-taxon 81-gene [supporting information (SI) Tables 1 and 2] data matrix with 76,583 aligned nucleotide positions, using maximum parsimony (MP), maximum likelihood (ML), and Bayesian inference (BI) methods. The MP analysis resulted in a single most parsimonious tree with a length of 165,426 steps, a consistency index (CI) of 0.357 (excluding uninformative characters), and a retention index (RI) of 0.593. All four ML analyses produced topologically identical trees (−lnL of 888368.804). BI analyses, using a single model for all genes (GTR + G + I) and a partitioned analysis employing five different models (SI Table 2), generated identical tree topologies with very similar posterior probabilities (PP) at each node. Each analysis resulted in one fully resolved tree (Fig. 1). Overall, support for monophyly of most clades was strong by all methods with BI and ML support generally higher than MP. Forty-eight of the 61 nodes had ≥95% bootstrap (BS) or ≥0.95 PP support values in trees using all three phylogenetic methods, and 10 additional nodes had ≥95% BS or ≥0.95 PP support values for two of three analyses.

All tree topologies based on nucleotide substitutions were identical with the exception that the MP tree differed from the ML and

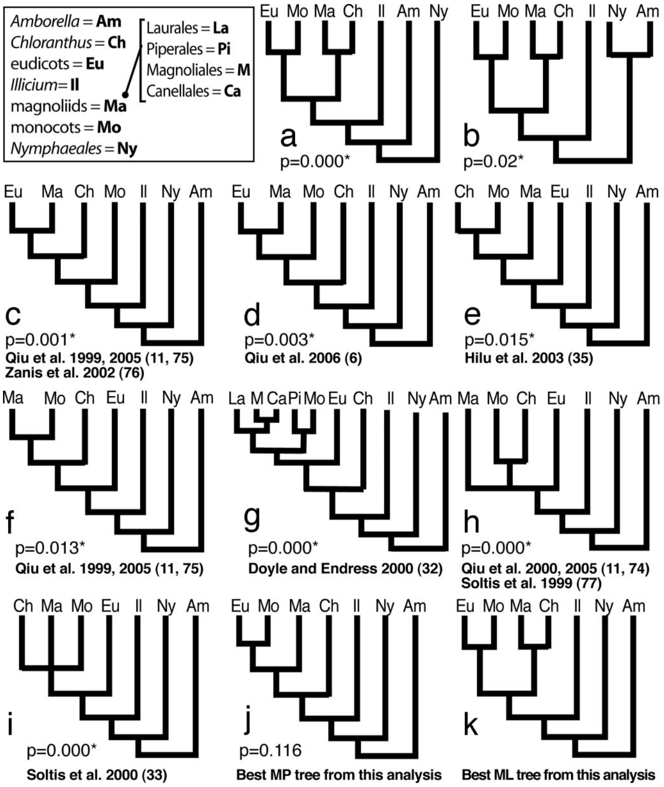


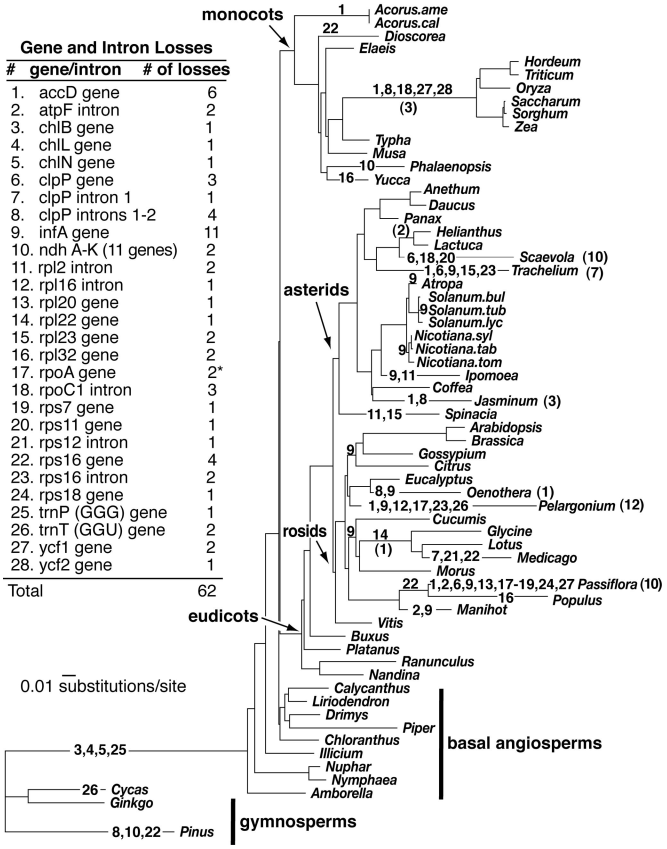
Fig. 2. Eleven alternative tree topologies used to perform the AU test for placement of *Amborella*, Nymphaeales, Chloranthales, magnoliids, monocots, and eudicots. (*a* and *b*) Modified topologies of the ML tree (*k*). (*c–i*) Topologies are derived from previous phylogenetic studies of angiosperms. (*j* and *k*) MP and ML/BI trees generated in this study.

BI trees in the placement of *Chloranthus* (Fig. 2 *j–k*). There was very strong support (100% BS, 0.99 PP) for the position of *Amborella* as the sole sister to the remaining angiosperms. The next diverging clade was Nymphaeales (*Nuphar* and *Nymphaea*) followed by Austrobaileyales (*Illicium*). In the ML and BI trees, Chloranthales (*Chloranthus*) were sister to a clade with four magnoliid orders, and this entire group was sister to a large clade that included both eudicots and monocots. Support for the sister relationship of Chloranthales and magnoliids was moderate (77% BS) or strong (1.0 PP) in ML or BI trees, respectively. In contrast, with MP Chloranthales were weakly supported (55% BS) as sister to a large clade of magnoliids, eudicots, and monocots.

MP analyses of a data matrix that included 1,268 indels, 226 of which were phylogenetically informative, yielded 75 MP trees of 1,305 steps (CI = 0.825, RI = 0.942). The strict consensus tree (data not shown) was congruent with but much less resolved than the MP tree based on nucleotides. The indels alone provided >50% BS support for the monophyly and relationships of 25 clades, and they increased the Bremer support values for 39 nodes (SI Fig. 4).

We used the approximately unbiased (AU) test to evaluate alternative tree topologies for relationships among the major angiosperm clades (Fig. 2). The ML tree (Fig. 2*k*) was found to be significantly better than 9 of 10 other trees (Fig. 2 *a–i*), including alternative resolutions of *Amborella*, Nymphaeales, magnoliids, Chloranthales, eudicots, and monocots. All tested hypotheses of relationships among these clades were rejected with $P < 0.05$, except for the topology generated in MP analyses in this study (Fig. 2*j*).

Changes in gene and intron content were plotted on the ML phylogram (Fig. 3). Most angiosperm plastid genomes contain 113 different genes, 16 of which are duplicated in the inverted repeat, for a total of 129 genes. This repertoire of genes is highly conserved among the basal lineages with all 62 losses confined to the more derived monocot and eudicot clades. Intron content is also highly conserved across angiosperms with most genomes containing 18 genes with introns. Like gene losses, intron losses are restricted to the more derived monocot and eudicot clades. There is a positive correlation between the number of base substitutions and indels ($r_s = 0.808$; $P \ll 0.001$; SI Fig. 5). There is also a positive correlation between branch lengths and the extent of genomic rearrangements (i.e., gene and intron losses and number of inversions, $r_s = 0.752$, $P \ll 0.001$; SI Fig. 6).



Gene and Intron Losses

#	gene/intron	# of losses
1.	accD gene	6
2.	atpF intron	2
3.	chlB gene	1
4.	chlL gene	1
5.	chlN gene	1
6.	clpP gene	3
7.	clpP intron 1	1
8.	clpP introns 1-2	4
9.	infA gene	11
10.	ndh A-K (11 genes)	2
11.	rpl2 intron	2
12.	rpl16 intron	1
13.	rpl20 gene	1
14.	rpl22 gene	1
15.	rpl23 gene	2
16.	rpl32 gene	2
17.	rpoA gene	2*
18.	rpoC1 intron	3
19.	rps7 gene	1
20.	rps11 gene	1
21.	rps12 intron	1
22.	rps16 gene	4
23.	rps16 intron	2
24.	rps18 gene	1
25.	trnP (GGG) gene	1
26.	trnT (GGU) gene	2
27.	ycf1 gene	2
28.	ycf2 gene	1
Total		62

Fig. 3. ML phylogram showing gene/intron losses and estimated numbers of inversions in angiosperm plastid genomes. Numbers in parentheses indicate the estimated number of inversions. The *rpl22* gene loss in *Gossypium* reported by Lee *et al.* (72) is not indicated, because this was an annotation error. *, *rpoA* has been reported missing from the *Pelargonium* plastid genome (73); however, a number of *rpoA*-like ORFs were identified. In *Passiflora*, a very divergent, potential pseudogene of *rpoA* is present based on the complete genome sequence (A.K.H., unpublished data). There is expression data that suggest that *rpoA* may be functional in *Pelargonium* (P. Kuhlman and J. D. Palmer, personal communication), but no expression data are available for *Passiflora*.

Discussion

Identification of the Earliest Diverging Angiosperm Lineage. Resolution of phylogenetic relationships among the basal clades of extant angiosperms is essential for addressing many important questions about the diversification and evolution of flowering plants. Despite intensive efforts during the past decade, considerable controversy persists concerning resolution of basal angiosperm relationships (4–8, 12, 13, 19, 21–25). Our phylogenetic analyses are the first to provide strong support for the position of *Amborella* as the sole sister group of the remaining angiosperms, using both MP and model-based analyses (Fig. 1). Resolution of this issue is clearly the result of the addition of both more taxa and more genes. Leebens-

Mack *et al.* (5) used 61 genes and 24 genomes, and MP trees placed *Amborella* alone sister to the remaining angiosperms, whereas ML analyses supported the *Amborella* + Nymphaeales basal hypothesis. The addition of 11 genomes for the same 61 genes (15) resulted in both MP and ML trees supporting *Amborella* alone as the first diverging lineage; however, support was weak in the ML tree, and SH tests could not reject the hypothesis of *Amborella* + Nymphaeales basal. Phylogenetic analyses of sequences of the 81 genes used here for these same 35 taxa (data not shown) provide strong support for *Amborella* as the earliest diverging lineage in both MP and ML trees, confirming that increasing the number of characters also contributes to the resolution of this issue. Strong support for the position of *Amborella* in the 81-gene, 64-taxon data set, combined with the rejection of alternative tree topologies (Fig. 2 *a* and *b*), provides convincing evidence that plastid genomic data unequivocally support *Amborella* as the sole sister group of the remaining angiosperms. Expanded sampling of genes and taxa from the mitochondrial and nuclear genomes is needed to confirm the plastid data, especially in view of the strong support of the *Amborella* + Nymphaeales hypothesis based on phylogenetic analyses of DNA sequences of three mitochondrial genes (6).

Relationships of Chloranthales, Magnoliids, Ceratophyllaceae, Monocots, and Eudicots. Despite intensive phylogenetic analyses of angiosperms, relationships among Chloranthales, magnoliids, Ceratophyllaceae, monocots, and eudicots remain especially controversial. Chloranthales, an ancient, isolated lineage dating back to the Early Cretaceous (26, 27), has been variously placed based on morphology and single or multiple gene sequences (Fig. 2 *c*–*i*), resulting in lack of an assignment of this order in the most recent angiosperm phylogeny group classification (3). The position of this order sister to magnoliids in both ML and BI trees and the rejection of seven of eight alternative topologies provides the strongest support so far for the placement of Chloranthales. Furthermore, recent phylogenetic analyses based on 17 plastid genes and one nuclear gene (*PHYC*) also provided moderate to strong support for a sister relationship between Chloranthales and magnoliids (9).

Placement of monocots has varied in analyses based on morphological characters or DNA sequences of single and multiple genes. Most studies placed monocots sister to either Ceratophyllaceae or magnoliids (28, 29), but with weak support. Previously, the highest level of support for a sister relationship between monocots and magnoliids occurred in a BI tree ($PP = 0.97$) based on four genes from plastid, nuclear, and mitochondrial genomes (29). In contrast, our phylogenetic analyses provide very strong support for the placement of monocots sister to eudicots (Fig. 1), and the AU test rejects topologies that include a sister relationship between monocots and magnoliids (Fig. 2*f*). A sister relationship between monocots and eudicots was also strongly supported in a recent phylogenetic analysis of DNA sequences of 17 plastid protein-coding genes and six associated noncoding regions (9), and Ceratophyllaceae were sister to eudicots with moderate support.

The recent completion of the *Ceratophyllum* plastid genome sequence (30) has provided new evidence for its placement among angiosperms. Phylogenetic analyses of 61 plastid genes for 45 taxa, using ML place *Ceratophyllum* sister to eudicots (Fig. 1 and supporting information figure 5 in ref. 30), although support for this relationship was only moderate in ML trees (71%). Furthermore, Moore *et al.* (30) were unable to reject many alternative topologies with the AU test (supporting information table 4 in ref. 30). In collaboration with Moore *et al.* (30), we added *Ceratophyllum* to our 64-taxon 81-gene data matrix, and results of both MP and ML analyses (SI Figs. 7 and 8) are congruent with their 61-gene 45-taxa data set; however, in the expanded matrix support of the placement of *Ceratophyllum* sister to eudicots in ML trees is stronger (82%; SI Fig. 8), and two thirds of the alternative topologies tested can be rejected in the AU test (SI Fig. 9).

Resolution of relationships among *Amborella*, Nymphaeales, Chloranthales, magnoliids, Ceratophyllaceae, monocots, and eudicots has very important implications for angiosperm evolution. The plastid genome tree provides an improved framework for comparative genomics in angiosperms and emphasizes the importance of including representatives from all of these clades. Current genomic sequencing projects only include representatives of the two most derived sister lineages, monocots and eudicots, but cDNA sequences for basal angiosperms are growing rapidly (31). An organismal context is essential for interpretation of nuclear gene phylogenies and the complex gene duplication history that they often imply. The phylogenetic framework also provides an improved platform for interpreting character evolution within angiosperms. The very rapid separation of monocots, eudicots, and magnoliids appears to have occurred in as few as 10 million years or less (5) and have given rise to >99% of angiosperms, but few if any morphological synapomorphies are known to resolve this critical event in angiosperm history (32).

Relationships of Major Eudicot Clades. Relationships among eudicot clades have been the focus of numerous phylogenetic studies during the past 15 years (reviewed in ref. 2). A consensus has been reached on the placement and circumscription of some clades, and, where those lineages are included, our results are congruent (Fig. 1), often with strong support. However, some eudicot relationships are controversial, and our analyses provide resolution of some of these outstanding issues.

Placement of Caryophyllales has been controversial, and phylogenetic analyses of single or multiple gene sequences have not been able to resolve their relationship to rosids or asterids (7, 33). Phylogenetic estimates based on 61 plastid protein-coding genes supported a sister relationship between Caryophyllales and asterids (5, 15, 16), although taxon sampling was limited in these studies. Our analyses with expanded taxon and gene sampling provide strong support for a sister relationship between asterids and *Spinacia*, the one member of Caryophyllales available. Expanded taxon sampling of this clade and other putative relatives is needed to confirm this relationship.

Relationships of Myrtales and Geraniales represent two of the remaining unsolved problems regarding deep-level relationships among rosids (2). They are currently treated as taxa of uncertain status due to their very inconsistent placement in molecular phylogenies based on one to several genes (7, 28, 34–36). Here, however, both orders are moderately or strongly supported as sister to the eurosid II clade in our tree based on 81 plastid genes (Fig. 1).

Evolution of Gene and Intron Content in Angiosperms. Gene and intron content are highly conserved among land plant plastid genomes, although losses have been identified in several angiosperm lineages (37, 38), and, in a few cases, there is evidence that genes have been transferred to the nucleus (*infA*, ref. 39; *rpl22*, ref. 40). Gene loss has been a common pattern throughout plastid genome evolution, presumably since the initial endosymbiotic event (41). Here, we have the opportunity to make the most extensive tree-based survey of the evolution of gene and intron content in seed plant plastid genomes (Fig. 3). The earliest diverging angiosperms include the complete repertoire of 129 genes, including 18 genes with introns. This pattern is conserved among major angiosperm clades, including Chloranthales, magnoliids, and the basal eudicots. Most of the 62 losses, involving 38 different genes and introns, are restricted to the more derived monocot and eudicot clades. Thus, it appears that plastid genomes of angiosperms were very stable during the period when many aspects of morphology, growth, and anatomy were undergoing extensive change. This pattern of stasis in gene and intron content of early diverging angiosperms and bursts of losses in multiple derived lineages parallels the situation in plant mitochondrial genomes (42).

Thirty genes have been lost from the 64 angiosperm and gymnosperm plastid genomes (Fig. 3). Four gene losses [*chlB*, *chlL*, *chlN*, and *trnP* (GGG)] represent synapomorphies for flowering plants. Three of these genes (*chlB*, *chlL*, and *chlN*) have been lost independently in two genera of Gnetales, *Gnetum* (43) and *Welwitschia* (44), but are present in *Ephedra* (L.A.R., unpublished data). Among angiosperms, six genes have been lost only once, in both our sampling and more extensive taxonomic surveys (37). The remaining 20 gene losses have occurred multiple times; 11 of these are represented by the loss of all *ndh* genes in both *Pinus* (45) and *Phalaenopsis* (46). The most common gene loss involves *infA*, with 11 independent losses in our sample (Fig. 3) and a minimum of 24 independent losses in angiosperms detected in an expanded survey (39). Also, *rps16* has been lost independently four times in our sample (Fig. 3), and expanded taxon sampling again provides evidence for more widespread loss of this gene throughout angiosperms (37). Likewise, of the eight introns that have been lost, five (*atpF*, *clpP* introns 1 and 2, *rpl2*, *rpoC1*, and *rps16*) have been lost multiple times. Expanded taxon sampling for intron losses has identified even more extensive convergent losses for *rpl2* (47), *rpl16* (37), *rpoC1* (48), and *rps12* (49). More intensive molecular investigations will likely reveal many more cases of gene and intron losses from the plastid genome and, in some cases, evidence for the transfer of these genes to the nuclear genome.

Correlation of Rates of Nucleotide Substitutions, Gene/Intron Losses, and Gene Order Changes in Plastid Genomes. Lineage-specific correlations between the numbers of base substitutions, indels, gene/intron content, and gene order changes were observed among angiosperm plastid genomes (Fig. 3 and SI Figs. 5 and 6). Elevated rates of nucleotide substitutions and indel evolution were detected on the branch leading to grasses (5). Our results confirm this earlier observation and suggest that lineage-specific rate accelerations may be a general feature of plastid genome evolution. Furthermore, we detected a positive correlation between increased substitution rates, numbers of gene/intron losses, and gene order changes. A similar phenomenon was documented in animal mitochondrial genomes (50, 51), but this is, to our knowledge, the first time such a correlation has been demonstrated in plastid genomes. The mitochondrial studies suggested several possible mechanisms to explain the correlation between increased rates of nucleotide substitution and genomic rearrangements, including efficiency of DNA repair, accuracy of DNA replication, metabolic rate, generation time, and body size. Xu *et al.* (51) argued that accuracy of DNA replication is the most likely explanation for increased rates of both nucleotide substitutions and genomic rearrangements in animal mitochondrial genomes.

More rigorous comparisons of rates of nucleotide substitutions, indels, and genomic rearrangements are needed to confirm the positive correlation between these changes and to explore mechanisms for this phenomenon. One possible explanation for this correlation may involve plastome mutator genes encoded in the nucleus. Experimental studies of mutants in *Oenothera* provided evidence that mutations in these genes can lead to enhanced rates of indels and base substitutions (52, 53). Furthermore, mutations of the *Arabidopsis* plastid mutator locus caused rearrangements in mitochondrial genomes (54), although we have no evidence supporting involvement of plastome mutator genes in any angiosperm lineages with accelerated rates of indels, base substitutions, or gene order changes. Expanded studies of these lineages, especially Campanulaceae, Geraniaceae, Goodeniaceae, and Passifloraceae, are needed to confirm this correlation and to explore possible mechanisms.

Materials and Methods

Taxon and Gene Sampling. The 64 taxa (SI Table 1) included here represent most major lineages of angiosperms (*sensu* APGII, 3), with three gymnosperms as outgroups. Angiosperm sampling included 41 different families from 33 orders. For 51 taxa, complete plastid genome sequences were available on GenBank, but 13 previously uncharacterized complete or draft genome sequences are reported here. Initially, DNA sequences of 83 genes were extracted from each genome, using DOGMA (55). Two of these genes (*ycf1* and *accD*) were later deleted because of alignment ambiguities resulting in a matrix of 77 protein-coding genes and four rRNAs (SI Table 2).

Plastid Isolation, Amplification, and Sequencing. Methods for isolating plastids and genome sequencing were described in refs. 56 and 57. Detailed protocols are available at www.jgi.doe.gov/sequencing/protocols/index.html.

Sequence Alignment. DNA sequences (SI Table 1) for 64 taxa were aligned by using a multiple sequence analysis tool (Z.C. and R.K.J., unpublished data). For protein-coding genes nucleotide sequences were translated into amino acids, aligned in MUSCLE (58) and manually adjusted. Nucleotide sequences were aligned by constraining them to the amino acid sequence alignment. A Nexus file was generated comprising 76,583 nucleotides (available at the Chloroplast Genome Database, http://chloroplast.cbio.psu.edu).

Phylogenetic Analyses. We estimated phylogenetic trees on the nucleotide substitution matrix, using MP (PAUP* software, Version 4.10; ref. 59), ML (GARLI software, Version 0.942; ref. 60), and BI (MrBayes software, Version 3.1.1; ref. 61). MP searches included 100 random addition replicates and TBR branch swapping with the Multrees option. Akaike information criterion via Modeltest software, Version 3.7 (62), was used to determine the most appropriate model for each of the 81 genes and for the full data matrix (GTR + G + I). We conducted four independent ML runs in GARLI, using the automated stopping criterion, terminating the search when the $-\ln$ score remained constant for 20,000 consecutive generations. Likelihood scores were calculated by using PAUP*, which better optimizes branch lengths (60). For BI, we performed two analyses: (*i*) all genes, GTR + G + I model; and (*ii*) genes partitioned into five different models (SI Table 2). Each run started with a random tree, default priors, and four Markov chains with heating values of 0.03, sampled every 100 generations. Analysis 1 ran for 6.3×10^7 generations, and analysis 2 ran for 8.0×10^7 generations. Convergence was confirmed by using AWTY graphical analysis (63). Fifty percent of burn-in trees were discarded. The frequency of inferred relationships was used to estimate PP. MP nonparametric BS analyses (64) were performed in PAUP with 1,000 replicates with TBR branch swapping, one random addition replicate, and the Multrees option and BS for ML analyses in GARLI with 100 replicates, using the automated stopping criterion set at 10,000 generations.

AU tests (65) conducted in Consel (66) were performed between the best ML tree and alternative topologies (Fig. 2 and SI Fig. 9) to evaluate whether likelihoods were significantly different.

Indels were coded by using simple indel coding (SIC, 67) and modified complex indel coding (MCIC, 68). Indels were included only when relative gap position in the alignment was entirely unambiguous. MCIC and SIC yielded similar results; therefore SIC coding, using SeqState (69), was used for all further analyses and yielded 1,268 indel characters, 226 of which were parsimony informative. The parsimony ratchet was used to find shortest trees based on indels only with 10 random addition runs of 200 ratchet cycles, using PRAP (70). Bremer support analysis was also performed with PRAP, using default settings. Bootstrapping was performed in PAUP (59), using 1,000 replicates, saving the 100 shortest trees per replicate.

Based on the MP topology inferred with nucleotides, ancestral indel character states were output for each node, using Mesquite (71), and parsed with a Perl script (written by K.F.M., available upon request) to determine the type of state transformations for each indel character on each branch, and to output data formatted for subsequent statistical analysis. Correlations of substitution rates (estimated from ML tree) with indels and with genome rearrangements were analyzed via Spearman's rank correlation coefficient and Spearman's *t* test.

We thank C. Morris (University of Connecticut, Hartford, CT) for supplying plant material; S. Plock for assistance with genome finishing; P. Piffanelli (Parco Tecnologico Padano, Lodi, Italy) for *Musa* and *Elaeis* plastid BACs; and A. Alverson, D. Hansen, J. D. Palmer, E. Ruck, and P. Steele for reading an earlier version of the manuscript. This work was supported by National Science Foundation Grants DEB 0120709 and DBI 0115684 (to R.K.J., L.A.R., C.W.d., and J.L.B.), National Institutes of Health Grant RO1 GM 063879, United States Department of Agriculture Grant 3611-21000-017-00D (to H.D.), and a Deutsche Telekom Stiftung scholarship (to K.F.M.).

1. Friis EM, Pedersen KR, Crane PR (2006) *Paleogeogr Paleoclimatol Paleoecol* 232:251–293.
2. Soltis PS, Endress PK, Chase MW, Soltis DE (2005) *Phylogeny and evolution of angiosperms* (Sinauer, Sunderland, MA).
3. Angiosperm Phylogeny Group (2003) *Bot J Linn Soc* 141:399–436.
4. Barkman TJ, Chenery G, McNeal JR, Lyons-Weiler J, Ellisens WJ, Moore G, Wolfe AD, dePamphilis CW (2000) *Proc Natl Acad Sci USA* 97:13166–13171.
5. Leebens-Mack J, Raubeson LA, Cui LY, Kuehl JV, Fourcade MH, Chumley TW, Boore JL, Jansen RK, dePamphilis CW (2005) *Mol Biol Evol* 22:1948–1963.
6. Qiu YL, Li L, Hendry T, Li R, Taylor DW, Issa MJ, Ronen AJ, Vekaria ML, White AM (2006) *Taxon* 55:837–856.
7. Soltis DE, Gitzendanner MA, Soltis PS (2007) *Int J Plant Sci* 168:137–157.
8. Graham SW, Olmstead RG (2000) *Am J Bot* 87:1712–1730.
9. Saarela JM, Rai HS, Doyle JA, Endress PK, Mathews S, Marchant AD, Briggs BG, Graham SW (2007) *Nature* 446:312–315.
10. Mathews S, Donoghue MJ (2000) *Int J Plant Sci* 161:S41–S55.
11. Qiu YL, Dombrovska O, Lee J, Li LB, Whitlock BA, Bernasconi-Quadroni F, Rest JS, Davis CC, Borsch T, Hilu KW, *et al.* (2005) *Int J Plant Sci* 166:815–842.
12. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2003) *Mol Biol Evol* 20:1499–1505.
13. Goremykin VV, Hirsch-Ernst KI, Wolfl S, Hellwig FH (2004) *Mol Biol Evol* 21:1445–1454.
14. Goremykin VV, Holland B, Hirsch-Ernst KI, Hellwig FH (2005) *Mol Biol Evol* 22:1813–1822.
15. Cai Z, Penaflor C, Kuehl JV, Leebens-Mack J, Carlson J, dePamphilis CW, Jansen RK (2006) *BMC Evol Biol* 6:77.
16. Hansen DR, Dastidar SG, Cai Z, Penaflor C, Kuehl JV, Boore JL, Jansen RK (2007) *Mol Phylogenet Evol* 45:547–563.
17. Felsenstein J (1978) *Syst Zool* 27:401–410.
18. Hendy MD, Penny DD (1989) *Syst Zool* 38:297–309.
19. Soltis DE, Soltis PS (2004) *Am J Bot* 91:997–1001.
20. Soltis DE, Albert VA, Savolainen V, Hilu K, Qiu YL, Chase MW, Farris JS, Stefanovic S, Rice DW, Palmer JD, Soltis PS (2004) *Trends Plants Sci* 9:477–483.
21. Stefanovic S, Rice DW, Palmer JD (2004) *BMC Evol Biol* 4:35.
22. Degtjareva GV, Samigullin TH, Sokoloff DD, Valiejo-Roman CM (2004) *Bot Zhurnal* 89:896–907.
23. Lockhart PJ, Penny D (2005) *Trends Plants Sci* 10:201–202.
24. Martin W, Deusch O, Stawski N, Grunhiet N, Goremykin VV (2005) *Trends Plants Sci* 10:203–209.
25. Goremykin VV, Hellwig FH (2006) *Gene* 381:81–91.
26. Eklund H, Doyle JA, Herendeen PS (2004) *Int J Plant Sci* 165:107–151.
27. Friis EM, Crane PR, Pedersen KR (1986) *Nature* 320:163–164.
28. Chase MW, Soltis D, Olmstead R, Morgan D, Les D, Mishler B, Duvall M, Price R, Hills H, Qiu YL, *et al.* (1993) *Ann Mo Bot Gard* 80:528–580.
29. Duvall MR, Mathews S, Mohammad N, Russell T (2006) *Aliso* 22:79–90.
30. Moore MJ, Bell CD, Soltis PS, Soltis DE (2007) *Proc Natl Acad Sci USA* 104:19363–19368.
31. Albert VA, Soltis DE, Carlson JE, Farmerie G, Wall PK, Ilut DC, Solow TM, Mueller LA, Landherr LL, Hu Y, *et al.* (2005) *BMC Plant Biol* 5:5.
32. Doyle JA, Endress PK (2000) *Int J Plant Sci* 161:S121–S153.
33. Soltis DE, Soltis PS, Chase MW, Mort MW, Albach DC, Zanis M, Savolainen V, Hahn WH, Hoot SB, Fay MF, *et al.* (2000) *Bot J Linn Soc* 133:381–461.
34. Savolainen V, Fay MF, Albach DC, Backlund A, van der Bank M, Cameron KM, Johnson SA, Lledo MD, Pintaud JC, Powell M, *et al.* (2000) *Kew Bull* 55:257–309.
35. Hilu KW, Borsch T, Müller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell M, Alice L, Evans R, *et al.* (2003) *Am J Bot* 90:1758–1776.
36. Savolainen V, Chase MW, Morton CM, Soltis DE, Bayer C, Fay MF, De Bruijn A, Sullivan S, Qiu YL (2000) *Syst Biol* 49:306–362.
37. Downie SR, Palmer JD (1992) in *Molecular Systematics of Plants*, eds Soltis PS, Soltis DE, Doyle JJ (Chapman & Hall, New York), pp 14–35.
38. Raubeson LA, Jansen RK (2005) in *Diversity and Evolution of Plants—Genotypic and Phenotypic Variation in Higher Plants*, ed Henry R (CABI, Wallingford, UK), pp 45–68.
39. Millen RS, Olmstead RG, Adams KL, Palmer JD, Lao NT, Heggie L, Kavanagh TA, Hibberd JM, Gray JC, Morden CW, *et al.* (2001) *Plant Cell* 13:645–658.
40. Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD (1991) *EMBO J* 10:3073–3078.
41. Martin M, Rujan T, Richly T, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) *Proc Natl Acad Sci USA* 99:12246–12251.
42. Adams KL, Qiu YL, Stoutemeyer M, Palmer JD (2002) *Proc Natl Acad Sci USA* 99:9905–9912.
43. Wu CS, Wang YN, Liu SM, Chaw SM (2007) *Mol Biol Evol* 24:1366–1379.
44. Burke DH, Raubeson LA, Alberti M, Hearst JE, Jordan ET, Kirch SA, Valinski AEC, Conant DS, Stein DB (1993) *Plant Syst Evol* 187:89–102.
45. Wakasugi T, Tsudzuki J, Ito T, Nakashima K, Tsudzuki T, Sugiura M (1994) *Proc Natl Acad Sci USA* 91:9794–9798.
46. Chang CC, Lin HC, Lin IP, Chow TY, Chen HH, Chen WH, Chen CH, Lin CY, Liu SM, Chang CC, Chaw SM (2006) *Mol Biol Evol* 23:279–291.
47. Downie SR, Olmstead RG, Zurawski G, Soltis DE, Soltis PS, Watson JC, Palmer JD (1991) *Evolution (Lawrence, Kans)* 45:1245–1259.
48. Downie SR, Llanas E, Katz-Downie DS (1996) *Syst Bot* 21:135–151.
49. McPherson MA, Fay MF, Chase MW, Graham SW (2004) *Syst Bot* 29:296–307.
50. Shao R, Dowton M, Murrell A, Barker SC (2003) *Mol Biol Evol* 20:1612–1619.
51. Xu W, Jameson D, Tang B, Higgs PG (2006) *J Mol Evol* 63:375–392.
52. Chang TL, Stoike LL, Zarka D, Schewe G, Chiu WL, Jarrell DC, Sears BB (1996) *Curr Genet* 30:522–530.
53. Stoike LL, Sears BB (1998) *Genetics* 149:347–353.
54. Martínez-Zapater JM, Gil P, Capel J, Somerville CR (1992) *Plant Cell* 4:889–899.
55. Wyman SK, Jansen RK, Boore JL (2004) *Bioinformatics* 20:3252–3255.
56. Jansen RK, Raubeson LA, Boore JL, dePamphilis CW, Chumley TW, Haberle RC, Wyman SK, Alverson AJ, Peery R, Herman SJ (2005) *Methods Enzym* 395:348–384.
57. McNeal JR, Leebens-Mack JH, Arumuganathan K, Kuehl JV, Boore JL, dePamphilis CW (2006) *BioTechniques* 41:69–73.
58. Edgar RC (2004) *BMC Bioinformatics* 5:1–19.
59. Swofford DL (2003) *PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods*, *Version 4.10* (Sinauer, Sunderland, MA).
60. Zwickl DJ (2006) *GARLI, Genetic Algorithm for Rapid Likelihood Inference, Version 0.942*. Available at www.bio.utexas.edu/faculty/antisense/garli/Garli.html.
61. Huelsenbeck JP, Ronquist F (2001) *Bioinformatics* 17:754–755.
62. Posada D, Crandall KA (1998) *Bioinformatics* 14:817–818.
63. Wilgenbusch JC, Warren DL, Swofford DL (2004) *AWTY*. Available at http://king2.scs.fsu.edu/CEBprojects/awty/awty_start.php.
64. Felsenstein J (1985) *Evolution (Lawrence, Kans)* 39:783–791.
65. Shimodaira H (2002) *Syst Biol* 51:492–508.
66. Shimodaira H, Hasegawa M (2001) *Bioinformatics* 17:1246–1247.
67. Simmons MP, Ochoterena H (2000) *Syst Biol* 49:369–381.
68. Müller KF (2006) *Mol Phylogenet Evol* 38:667–676.
69. Müller KF (2005) *Appl Bioinformatics* 4:65–69.
70. Müller KF (2004) *Mol Phylogenet Evol* 31:780–782.
71. Maddison WP, Maddison DR (2006) *Mesquite, Version 1.12*. Available at http://mesquiteproject.org.
72. Lee SB, Kaittanis C, Jansen RK, Hostetler JB, Tallon LJ, Town CD, Daniell H (2006) *BMC Genomics* 7:61.
73. Chumley TW, Palmer JD, Mower JP, Fourcade HM, Calie PJ, Boore JL, Jansen RK (2006) *Mol Biol Evol* 23:2175–2190.
74. Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW (2000) *Int J Plant Sci* 161:S3–S27.
75. Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen ZD, Savolainen V, Chase MW (1999) *Nature* 402:404–407.
76. Zanis MJ, Soltis DE, Soltis PS, Mathews S, Donoghue MJ (2002) *Proc Natl Acad Sci USA* 99:6848–6853.
77. Soltis PS, Soltis DE, Chase MW (1999) *Nature* 402:402–404.