LBNL-58174

Identifying the Basal Angiosperm Node in Chloroplast Genome Phylogenies: Sampling One's Way Out of the Felsenstein Zone

Jim Leebens-Mack[*], Linda A. Raubeson[¶], Liying Cui[*], Jennifer V. Kuehl[‡], Matthew H. Fourcade[‡], Timothy W. Chumley[§], Jeffrey L. Boore[‡†], Robert K. Jansen[§], and Claude W. dePamphilis[*]

[*]Department of Biology, Institute of Molecular Evolutionary Genetics, and The Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802; [¶]Department of Biological Sciences, Central Washington University, Ellensburg, WA 98926; [‡]DOE Joint Genome Institute, Walnut Creek, CA 94598; [§]Section of Integrative Biology and Institute of Cellular and Molecular Biology, University of Texas, Austin TX 78712; [†]Department of Integrative Biology, University of California, Berkeley CA 94720.

Research Article

Key words: Angiosperm phylogeny, phylogenomics, parametric bootstrap.

Running Head: Chloroplast Genome Phylogenies

Corresponding author: Jim Leebens-Mack, Department of Biology, Institute of Molecular Evolutionary Genetics, and The Huck Institutes of Life Sciences, The Pennsylvania State University, 402 Life Sciences Building, University Park, PA 16802; phone:814-865-3083; FAX: 814-863-1357; email: jleebensmack@psu.edu

Abstract

While there has been strong support for *Amborella* and Nymphaeales (water lilies) as branching from basal-most nodes in the angiosperm phylogeny, this hypothesis has recently been challenged by phylogenetic analyses of 61 protein-coding genes extracted from the chloroplast genome sequences of *Amborella*, *Nymphaea* and 12 other available land plant chloroplast genomes. These character-rich analyses placed the monocots, represented by three grasses (Poaceae), as sister to all other extant angiosperm lineages. We have extracted protein-coding regions from draft sequences for six additional chloroplast genomes to test whether this surprising result could be an artifact of long-branch attraction due to limited taxon sampling. The added taxa include three monocots (*Acorus, Yucca* and *Typha)*, a water lily (*Nuphar)*, a ranunculid (*Ranunculus)*, and a gymnosperm (*Ginkgo*). Phylogenetic analyses of the expanded DNA and protein datasets together with microstructural characters (indels) provided unambiguous support for *Amborella* and the Nymphaeales as branching from the basal-most nodes in the angiosperm phylogeny. However, their relative positions proved to be dependent on method of analysis, with parsimony favoring *Amborella* as sister to all other angiosperms, and maximum likelihood and neighbor-joining methods favoring an *Amborella* + Nympheales clade as sister. The maximum likelihood phylogeny supported the later hypothesis, but the likelihood for the former hypothesis was not significantly different. Parametric bootstrap analysis, single gene phylogenies, estimated divergence dates and conflicting indel characters all help to illuminate the nature of the conflict in resolution of the most basal nodes in the angiosperm phylogeny. Molecular dating analyses provided median age estimates of 161mya for the most recent common ancestor of all extant angiosperms and 145 mya for the most recent common ancestor of monocots, magnoliids and eudicots. Whereas long sequences reduce variance in branch lengths and molecular dating estimates, the impact of improved taxon sampling on the rooting of the angiosperm phylogeny together with the results of parametric bootstrap analyses demonstrate how long-branch attraction can mislead genome-scale phylogenetic analyses.

Introduction

Characterized by Darwin (1903) as "an abominable mystery", the early radiation of angiosperms was pivotal in the evolutionary history of our biota. After years of controversy concerning identity of the basal-most node in the angiosperm phylogeny, a series of studies using multiple genes from the chloroplast, mitochondrial and nuclear genomes identified *Amborella*, the Nymphaeales, and Austrobaileyales as successive sister lineages relative to all other angiosperms (Mathews and Donoghue 1999; Qiu et al. 1999; Soltis, Soltis and Chase 1999; Parkinson, Adams and Palmer 1999; Graham and Olmstead 2000; Barkman et al. 2000; Zanis et al. 2002; Borsch et al. 2003; Hilu et al. 2004; Stefanović, Rice and Palmer 2004; e.g., fig. 1A). Many of these studies found that the inferred relationship between *Amborella* and the Nymphaeales varied when using differing methods of phylogenetic reconstruction, models of molecular evolution, and subsets of taxa (Barkman et al. 2000; Graham and Olmstead 2000; Zanis et al. 2002; Stefanović, Rice and Palmer 2004), with each lineage sometimes inferred to be most basal and, in some cases, for the two to form a single clade, sister to all other angiosperm lineages (e.g., fig. 1B, C). While the branching order of *Amborella* and the Nymphaeales relative to each other and the rest of the angiosperms has remained controversial, there has been widespread consensus in the recent plant systematics literature for *Amborella* and Nymphaeales branching off at the base of the angiosperm phylogeny, followed subsequently by Austrobaileyales and the remaining angiosperm lineages.

That consensus was recently challenged by the results of phylogenetic analyses of 61 protein-coding genes common to 14 chloroplast genome sequences including the recently sequenced plastid genomes of *Amborella trichopoda* (Goremykin et al. 2003) and *Nymphaea alba* (Goremykin et al. 2004). The analyses of Goremykin and colleagues placed the monocots, represented by the chloroplast genomes of rice (*Oryza sativa*), maize (*Zea mays*) and wheat (*Triticum aestivum*), as sister to all other extant lineages in the angiosperm phylogeny. This result is quite intriguing because, much of our current understanding of morphological, developmental and molecular evolution in early angiosperm history would have to change if monocots are in fact sister to all other angiosperms (e.g. fig 1D). As Goremykin et al. (2004; p. 1452) point our in their second paper, however, the hypothesized basal divergence of monocots from all other angiosperms must be tested in analyses with increased taxon sampling. Genome-scale phylogenetic studies, where character sampling is very deep but taxon sampling is typically sparse, are particularly susceptible to long branch attraction (Soltis et al. 2004a; Philippe, Lartillot and Brinkmann 2005), and multiple lines of evidence suggest that the placement of the grasses in the analyses of Goremykin et al. (2003; 2004) may be an artifact of sparse taxon sampling, particularly within the monocots (Soltis and Soltis 2004a; Soltis et al. 2004a; Stefanović, Rice and Palmer 2004; but see Lockhart and Penny 2005; Martin et al. 2005). Here we test this hypothesis directly by adding the corresponding 61 gene datasets for six additional species, including three non-grass monocot species, a ranunculid (the sister clade to all other eudicots), an additional water lily and the gymnosperm *Ginkgo biloba*. Our study also includes an evaluation of microstructural mutations, a class of evolutionary events that may be less subject to homoplastic events than base substitutions. In addition, we extend the earlier work of Zanis et al. (2002) using parametric bootstrap analyses to investigate estimated branching events at the base of the angiosperms and estimate minimum divergence dates for well supported clades on the phylogeny.

Materials and Methods

Sequencing

Templates suitable for constructing random insert plasmid libraries were generated for the chloroplast genomes of *Ginkgo biloba, Nuphar advena*, *Acorus americanus*, *Yucca schidigera*, *Typha latifolia*, and *Ranunculus macranthus* by one of three ways (Jansen et al. 2005): (1) sucrose gradient isolation of pure cpDNA (*Ginkgo*, *Nuphar* and *Ranunculus*); (2) rolling circular amplification (RCA) of the entire plastid genome using crude cpDNA isolations

(*Acorus* and *Typha*); and (3) cpDNA-containing clones identified through screening of a fosmid genomic library (*Yucca*). The low coverage (<0.5x) yucca fosmid library was produced using the Epicentre CopyControl™ Fosmid Library Production Kit according to manufacturer protocols (http://www.epicentre.com). All templates were sheared by serial passage through a narrow aperture using a Hydroshear® device (Genomic Solutions, Ann Arbor, MI) into random fragments of approximately 3 kb, then cloned into the pUC18 plasmid vector to create a clone library. These clones were robotically processed through colony picking, rolling circle amplification using TempliPhi™ (Amersham Biosciences), sequencing reactions with either ET terminators (Amersham Biosciences) or BigDye™ terminators (Applied Biosystems), then reads of approximately 700 nucleotides each were determined from each end of each clone. Detailed protocols are available at http://www.jgi.doe.gov/sequencing/protocols/prots_production.html. Sequence reads were trimmed and assembled using phred and phrap (Ewing et al. 1998; Ewing and Green, 1998) and manually interpreted using Consed (Gordon, Abajian and Green, 1998) and Sequencher (Gene Codes Corp, 2003). Roughly 4,000 sequencing reads were determined for each chloroplast genome, giving about 2.8 million nucleotides of sequence (at Q20 or greater quality), for an average depth of coverage in the assemblies of about 12 X (considering losses due to impure cpDNA preparations). The sequence of each cpDNA was finished to meet the quality criteria specified in Jansen et al. (2005) by additional sequencing reactions that targeted gaps or parts of the genomes with potential errors using primers specifically designed to portions of the plastid genome.

Phylogenetic Analyses

The 61 genes included in the analyses of Goremykin et al. (2003, 2004) were extracted from high quality contigs (Q values > 40 for all extracted bases) of our six new chloroplast genome sequences using the organellar genome annotation program DOGMA (http://evogen.jgi-psf.org; Wyman, Jansen and Boore 2004). The same set of 61 genes was extracted from chloroplast genome sequences for the 18 other available species (table 1). Inferred amino acid sequences for each of the 61 genes were aligned using CLUSTAL W (Thompson et al. 1994) and adjusted manually. A nucleotide alignment was then forced to correspond to the amino acid alignment and further adjusted. The complete amino acid and nucleotide alignments are available in our Chloroplast Genome Database (http://chloroplast.cbio.psu.edu/) and sequences are available in GenBank (see table 1 for accession numbers).

Phylogenetic analyses of nucleotide alignments using maximum parsimony (MP) and neighbor-joining (NJ) were performed using PAUP* version 4.10 (Swofford 2003), and maximum likelihood (ML) analyses were performed using both PAUP* and PHYML v. 2.4.4 (Guindon and Gascuel 2003). All MP searches were heuristic with 10 random addition replicates and TBR branch swapping. The Hasegawa-Kishino-Yano (HKY; 1985) model of molecular evolution was used in ML and NJ analyses of the nucleotide alignments. ML estimates of HKY distances were used for the NJ analyses. ML and NJ analyses of amino acid alignments were performed using PHYML and PHYLIP v. 3.63 (seqboot, protdist and neighbor; Felsenstein 2004), respectively, under the Jones-Taylor-Thornton model (JTT; 1992). ML and NJ analyses of the nucleotide alignment were run with and without rate variation among sites (HKY+Γ), with and without invariant sites (HKY+Γ+I). ML and NJ analyses of the amino acid alignment were run with and without rate variation among sites (JTT+Γ), and ML analyses were also run with and without invariant sites (JTT+Γ+I). Rate variation among sites was estimated as a discrete gamma distribution with six rate classes (Yang 1994). ML parameter estimates for rate variation across sites and for invariant sites were optimized simultaneously with topology and branch-lengths in PHYML. In order to avoid being trapped in local optima, all ML analyses performed in PHYML were run with five starting trees including the BIONJ tree (default) and the four trees shown in figure 1. Likelihood ratio tests showed significant improvement in the fits of both nucleotide and protein evolution models with the addition of parameters for rate variation across sites and invariant sites (p << 0.001). Non-parametric bootstrap analyses (Felsenstein 1985) were performed for all analyses with 200 pseudoreplicates.

In addition to the typical ML analysis, constrained likelihood trees were estimated for the nucleotide alignment setting one of four lineages as sister to the remaining angiosperms: *Amborella* (fig. 1A) *Amborella* + Nymphaeales (fig. 1B), Nymphaeales (fig. 1C) and monocots (fig. 1D; Goremykin et al 2003; 2004). The Shimodaira-Hasegawa (SH; 1999) test was performed in PAUP* to determine whether any of the resulting phylogenies was significantly worse than the ML tree. The test was performed using 10,000 bootstrap replicates. MP topologies were tested similarly in PAUP* using the Wilcoxon test.

In their analyses, Goremykin et al. (2003; 2004) removed third codon positions from their nucleotide alignments citing concerns over saturation of nucleotide substitutions at synonymous sites. Pair-wise genetic distances for all taxon pairs in our analysis (HKY model) were calculated separately for codon positions 1+2 and position 3. A linear relationship was observed between genetic divergence (HKY distances) at third versus first and second positions (fig. 2) suggesting that saturation was not seriously biasing distance estimates, so third positions were included in our phylogenetic analyses. The inclusion of third codon positions changed bootstrap values for some nodes, but with the exception of the poorly supported placement of *Calycanthus* the optimal MP, ML and NJ topologies were the same in 61-gene analyses with and without their inclusion (see results).

Initial analyses of codon usage, amino acid content and branch-lengths showed that the fern *Adiantum* and hornwort *Anthoceros* sequences are extremely divergent from their closest relatives in the study, in part due to extensive RNA editing (Kugita et al. 2003a; Wolf et al. 2003). These taxa, which were not critical to our study, were eliminated from further analyses. Gapped sites were also excluded from additional analyses because these often represented regions of questionable annotation and alignment. An exception was made for well-aligned sites in *rpoA* and *ccsA*, which were missing and coded as gaps in *Physcomitrella patens*. The resulting nucleotide and amino acid alignments included 39,978 and 13,326 characters, respectively. Exclusion of *Adiantum*, *Anthocerous* or gapped sites affected only the poorly supported placement of *Calycanthus* relative to the eudicots in MP, NJ, and ML topologies (data not shown).

Ané et al. (2005) have recently shown among-lineage rate heterogeneity, heterotachy (Lopez, Casane and Philippe 2002) at many sites in plant plastid genes. This type of rate variation can confound phylogenetic analyses given some patterns of variation (e.g. Kolaczkowski and Thornton 2004; Spencer Susko and Roger 2005). We attempted to control for heterotachy in distance analyses of both nucleotide and amino acid alignments by estimating LogDet distances (Steel 1994; Lockhart et al. 1994, Lake 1994) with and without among-site rate heterogeneity using LDDist (Thollesson 2004; Martin et al. 2005).

Parsimony analyses were also performed on all 61 genes individually. Analyses were performed on alignments of all three codon positions, first and second codon positions and amino acids. Gapped positions were not included the analyses. Bootstrap analyses were performed with 250 replicates and all of the resulting phylogenies were inspected to identify the most basal angiosperm lineage in cases where one clade was identified as sister to all other angiosperms with at least 50% bootstrap support.

Whereas alignment in many gapped regions was problematic, there were many regions where the homology of insertions or deletions could be assigned unambiguously. A separate data matrix of insertions and deletions was constructed from these regions and parsimony analysis was performed on the binary data matrix.

Parametric Bootstrap Analyses

While much has been made of *Amborella* as the most basal clade in the angiosperm phylogeny (Mathews and Donoghue 1999; Qiu et al. 1999; Soltis, Soltis and Chase 1999; Parkinson, Adams and Palmer 1999; Barkman et al. 2000; Graham and Olmstead 2000; Zanis et al. 2002; Borsch et al. 2003; Hilu et al. 2003), some analyses in these studies and others have found support for the placement of *Amborella* and Nymphaeales in a clade placed sister to all other angiosperms whereas others have shown evidence for Nymphaeales alone as the most basal clade in the angiosperm phylogeny. Barkman et al. (2000) first showed that the placement of *Amborella* relative to Nymphaeales could vary among datasets and methods of phylogenetic reconstruction and predicted that incongruence among phylogenies estimated using different methods would increase as additional sequence data were added for each taxon in genome-scale analyses due to the effects of long branch attraction (figure 5 of Barkman et al. 2000). Following the comprehensive investigation of Zanis et al. (2002), we used parametric bootstrap analyses to explore whether long branch attraction might be responsible for the observed conflict and determine if any of the four hypotheses illustrated in figure 1 could be rejected given the large number of chloroplast genome sequences in our study.

ML phylogenies were estimated in PAUP* with constraint trees corresponding to one of four hypothesized lineages as sister to all other extant angiosperms: **a**) *Amborella*, **b**) an *Amborella* + Nymphaeales clade, **c**) Nymphaeales, and **d**) monocots (fig. 1). Parameter values for the HKY model with invariant sites and among-site rate variation were estimated in PHYML as described above. These parameter values and each of the four ML phylogenies shown in figure 1 were used to generate 200 simulated datasets with Seq-Gen v1.3 (Rambaut and Grassly1997). MP and NJ analyses were performed on all simulated datasets in PAUP* and ML analyses were performed using PHYML with five starting trees as described above. Model parameters for the ML and NJ analyses matched those used to simulate the data. The frequencies with which hypothesis **a**, **b**, **c** or **d** were observed in the estimated phylogenies was calculated for each set of simulations and each phylogenetic methodology.

Dating Nodes

If some speciation events that gave rise to basal angiosperm lineages were separated by just a few million years, it may be difficult to resolve these events even with the large amount of data included in genome-scale studies with adequate taxon sampling. In order to explore timing of major events in angiosperm history, the ML phylogenies generated for each of 200 non-parametric replicate were saved with branch lengths estimated to eight significant digits and minimum ages were estimated for nodes on each phylogeny using the penalized likelihood method implemented in r8s (Sanderson 2003). The origin of the eudicots 125 mya as evidenced by the apperence of tricolpate pollen in the fossil record (Crane, Friis and Pedersen 1995; Sanderson et al. 2004) was used as fixed calibration point. In addition, minimum and maximum ages for the origin of the euphyllophytes were set at 380 and 410 mya, respectively (Schneider et al. 2004), and minimum ages for the Poaceae and the most recent common ancestor of *Ginkgo* and *Pinus* were 55 mya (Kellogg 2000) and 310 mya (Schneider et al. 2004), respectively. Averages and standard errors were estimated for the most recent common ancestor defined by the nodes in the ML tree across all 200 bootstrap trees.

Results

Monocots are not sister to all other extant angiosperms

As Felsenstein (1978) famously deduced, there is a tendency for long branches to artifactually attract (latter referred to as "long branch attraction") in some phylogenetic analyses, and under these conditions, the probability of inferring the wrong phylogeny will increase with additional data due to statistical inconsistency. Hendy and Penny (1989) showed that even when rates of evolution are constant among lineages, long-branch attraction may confound parsimony analyses with more than four taxa and trees with variable terminal branch lengths. Since the outgroup is almost always a long branch, this often manifests as the longer branch ingroup taxa being drawn to the base of the tree even when this is not the correct relationship (Philippe and Laurent 1998). In such instances, adding taxa to interrupt the longest branches can help parsimony to converge on the correct phylogeny (Hendy and Penny 1989).

By adding non-grass monocots and another gymnosperm to the analysis, we interrupted two of the longest branches (fig. 3). As a result, all of MP and ML analyses placed branching points for *Amborella* and Nymphaeales at the base of the angiosperm phylogeny with strong support (fig. 4). The most parsimonious nucleotide-based phylogeny constrained to place monocots sister to all other angiosperms was 242 steps longer than the unconstrained MP phylogeny (fig. 4B). A Wilcoxon signed ranks test showed this difference to be highly significant ($p < 0.0001$). ML analyses with among-site rate variation placed *Amborella* + Nymphaeales sister to all other angiosperms (fig. 4A), whereas the ML analysis without rate variation placed *Amborella* as sister to all extant angiosperms (supplemental fig. 1A and 1B). The log-likelihood score for the optimal HKY-Γ+I likelihood tree in an analysis with constrained to place monocots sister to all other angiosperms (-ln(L) = 318423.61) was 310.73 points greater than the score for the unconstrained ML tree (-ln(L) = 318112.88; figs. 1B and 4A). Based on the Shimodaira-Hasegawa (S-H; 1999) test, this difference in likelihood scores was highly significant ($p < 0.0001$). The NJ analyses with among-site rate variation found strong support for an *Amborella* + Nymphaeales clade sister to all other angiosperms (fig. 4C), but the NJ analysis under the simpler model without variation in rates among sites placed the eudicots as sister to the remaining angiosperms (supplemental fig. 1C). NJ analyses with the simple HKY model performed after removing the most distant outgroup taxa, *Physcomitrella*, *Marchantia* and *Psilotum*, placed an *Amborella* + water lilies clade as sister to the remaining angiosperms (supplemental fig. 1D). ML and NJ analyses performed on the nucleotide data matrix using the more complicated GTR-Γ+I model gave the same topologies shown in figure 4 with bootstrap values ± 2% relative to the results of the HKY-Γ+I analysis (data not shown).

Analyses of the amino acids and first and second codon positions also placed *Amborella* and Nymphaeales at the base of the angiosperm phylogeny with strong support in the MP and ML analyses (supplemental figs. 2 and 3, respectively). NJ analyses of the first and second positions gave the same topology as the 3 codon position analysis, but the amino acid analysis placed eudicots as sister to the remaining angiosperms giving weak support for a (monocots (*Calycanthus*, (*Amborella*, Nymphaeales))) clade (supplemental figs. 2 and 3, respectively). The placement of *Calycanthus,* which was poorly supported in all but two NJ analyses, differed among all three phylogenetic methods in nucleotide analyses (fig. 4). Whereas bootstrap support for the placement of *Calycanthus* was ≤ 65 % in all MP and ML analyses, *Calycanthus* was placed sister to the monocots with support values of 100% and 90% in NJ analyses of the ungapped nucleotide and first and second codon position alignments, respectively. Despite high support values in these NJ analyses, additional sequences sampled from magnoliid orders other than Laurales (Magnoliales, Piperales, and Canellales), a member of the Chloranthaceae, and *Ceratophyllum* will be necessary to have any hope of resolving the relationships among these taxa, the monocots and eudicots, with confidence.

NJ analyses of LogDet distance matrices estimated from nucleotide and amino acid alignments assuming no among-site rate variation, six rate classes, and invariant sites (ML estimate of 26%) plus six additional rate classes all produced the same topology placing the core eudicots sister to a clade with all other eudicots including the eudicot *Ranunculus* (supplemental fig. 4). The monophyly of the eudicots including the ranunculids as sister to the core eudicots is well established (Judd and Olmstead 2004) so this topology is quite unlikely to represent the true phylogeny. Moreover, analyses of nucleotide alignments including only seed plants, gave strong support for an *Amborella* + water lilies clade as sister to the remaining angiosperms (supplemental fig. 4 E and F). The amino acid analysis of the seed plants, however, placed eudicots (including *Ranunculus*) sister to the remaing angiosperms (supplemental fig. 4 G and H). These result is noteworthy as they add to a growing set of examples underscoring the need for deeper understanding of how covarion/covariotide evolution and other forms of heterotachy can be diagnosed and modeled in phylogenetic analyses (e.g. Lockhart et al. 1998; 1999; Lopez, Casane and Philippe 2002; Huelsenbeck 2002; Phillips, Delsuc, and Penny 2004; Ané et al. 2005; Kolaczkowski and Thornton 2004; Spencer Susko and Roger 2005; Martin et al. 2005).

In general, the relationships common to all trees in figure 4 were found all ML trees and most NJ trees inferred from nucleotide and amino acid alignments given a variety of substitution models. Most of the exceptions seen in the NJ trees were only found in analyses that included distantly related outgroup taxa. Most importantly, conflicts among topologies derived using different phlogenetic methods, substitution models or taxon sets were usually supported by bootstrap values greater than 90%. This observation underscores how statistical inconsistency is a serious problem for phylogenomics that can only be diagnosed through comprehensive analyses using a variety of phylogenetic methods, substitution models, distance corrections and taxon sets (see also Phillips, Delsuc, and Penny 2004).

Amborella or Amborella plus Nymphaeales?

As has been described in previous studies (Barkman et al. 2000; Zanis et al. 2002; Stefanović et al. 2004), the relationship between *Amborella* and the Nymphaeales relative to the rest of the angiosperms depended on the method of analysis. In our MP analyses, *Amborella* was placed sister to all other extant angiosperms with high support (fig. 4B). A tree constrained to have a Nymphaeales + *Amborella* clade was 83 steps longer than the MP tree and this difference was found to be significant in a Wilcoxon signed-ranks test (p = 0.0007). Examination of the characters potentially supporting an *Amborella* alone vs. a Nymphaeales + *Amborella* clade showed that characters were distributed evenly across the 61 gene alignment. However, despite strong support for the basal branching point for *Amborella* under MP, the NJ and ML analyses unite Nymphaeales with *Amborella* in a clade that is sister to the rest of the angiosperms. The ML analysis gave moderate support for this relationship (fig. 4A) whereas the NJ analysis provided strong support (fig. 3C). The likelihood score (-ln(L) = 318,117.04) for a tree constrained to place *Amborella* as sister to all other angiosperms was not significantly different than that of the ML tree (p = 0.601; scores shown in fig. 1). Furthermore, ML analyses the amino acid alignments provided weak support for *Amborella* as sister to the remaining angiosperms (supplemental fig. 3). The tree placing the water lilies as sister to the rest of the angiosperms gave a significantly worse likelihood score than the ML phylogeny (p = 0.028).

Most single gene analyses did not resolve basal relationships among angiosperm lineages (table 2). In all character sets, however, when one angiosperm lineage was placed sister to the others, it was most often *Amborella* or *Amborella* + Nymphaeales (table 2). The monocots were not resolved as sister to the other angiosperms in any of the single gene phylogenies, and Poaceae or Poales (Poaceae + *Typha*) were estimated as the most basal clade for only four genes in analyses of codon positions 1 and 2. A few of the single gene analyses supported relationships that are clearly untenable (table 2). For example, a weakly supported Poaceae + *Oenothera* clade was placed sister to the rest of the angiosperms in the analysis of *clpP* nucleotides. The most basal angiosperm lineage inferred for some genes varied across the three character sets. For example, the *psbB* phylogenies estimated from all nucleotides, first and second codon positions and amino acids were core eudicots, Poales, and *Typha*, respectively.

The MP analysis of insertions and deletions also provided strong support for *Amborella* and Nymphaeales as the most basal angiosperm lineages (fig. 5 and supplemental tables 1, 2). A total of 116 potentially phylogenetically informative indels were included in the analysis. Four unambiguous indels, consisting of three insertions and one deletion mutation supported the monophyly of all sampled angiosperms except *Amborella* and Nymphaeales (fig. 6). No indel characters were potentially supportive of grasses or monocots in the basal most position. Twelve equally parsimonious trees (141 steps) were recovered in the MP analysis of the indel characters. Half of the MP trees included an *Amborella* + Nymphaeales clade and half placed *Amborella* alone as sister to all other angiosperm lineages. Each of these two hypotheses was supported by a single synapomorphy (fig. 6). The resolved portions of the indel phylogeny were identical to corresponding relationships inferred through comparisons of nucleotide and amino acid sequences.

A large number of the microstructural mutations distinguish grasses from all other plants in the study (fig. 5B), whereas other lineages have many fewer insertion-deletion mutations. This suggests that lineage-specific common processes may have given rise to the enhanced rates of nucleotide substitution and indel evolution in the lineage leading to the Poaceae, and perhaps generally throughout the major lineages of land plant evolution.

Parametric Bootstrap Analysis

The parametric bootstrap analysis demonstrated that the results of the MP analysis may have been affected by long branch attraction (Felsenstein 1978; Hendy and Penny 1989). Whereas the ML and NJ analyses recovered the simulated topology in the vast majority of cases, the MP analyses misidentified *Amborella* as sister to all other angiosperms in 21% of the cases where data were simulated under the *Amborella* + Nymphaeales phylogeny (table 3). When constrained to place monocots sister to the rest of the angiosperms, the ML phylogeny included a very short internal branch leading to the node where an *Amborella* + Nymphaeales clade diverged from the remaining dicot lineages (fig 1D). Datasets generated under this hypothesis were especially problematic for MP and NJ analyses. Only ML recovered the correct tree in more than half of the analyses (table 3). In exploratory parametric bootstrap analyses where Nymphaeales was only represented by the *Nuphar* sequence, both NJ and MP analyses failed to recover the correct topology for more than 97% of the data sets simulated under the *Amborella* + Nymphaeales phylogeny. The effect of adding *Nymphaea* to the analysis illustrates the strong influence of taxon sampling on phylogenetic reconstruction.

Molecular Clock Estimates

Molecular clock estimates for most angiosperm nodes in the ML topology were in line with recently published divergence dates estimated using a variety of procedures (Chaw et al. 2004; Davies et al. 2004; others reviewed in Sanderson et al. 2004; table 4; fig. 7). The median age estimates for the angiosperm crown group and the most recent common ancestor of all monocots, magnoliids and eudicots were 161 mya and 145 mya, respectively. The bootstrap distribution of divergence time estimates was unimodal with low standard error estimates for most nodes with a notable exception (table 4). Age estimates for the most recent common ancestor of *Amborella* and the Nymphaeales had two modes, corresponding to bootstrap replicate

ML topologies with and without an *Amborella* + Nympheaeles clade. Trees with *Amborella* and the water lilies forming a clade gave a median age estimate of 135 mya for their most recent common ancestor (MRCA), whereas the median age estimate was 161 mya for the 68 bootstrap replicates where *Amborella* was found to be sister to all other angiosperms.

The range of age estimates calculated from 200 bootstrap replicates is shown in table 4. It should be noted, however, that whereas the small errors in our branch-lengths estimates follow legitimately from the large number of nucleotide positions included in our analyses, the errors on our divergence date estimates are artificially low given that our dating analysis was done under the assumption that calibration points for the most recent common ancestors of all eudicots is known without error (see Graur and Martin 2004). The appearance of tricolpate pollen in the fossil record at the Barremian-Aptian boundary 125 mya provides a minimum age for the most recent common ancestor of all extant eudicots.

Discussion

Phylogenetic analyses of plant plastid genomes are providing new insights into the evolution of gene order (refs) lineage-specific substitution rates and patterns (Ané et al. 2005) and factors influencing genome-scale phylogenetic inference (Lockhart et al 1999; Goremykin et al 2003; 2004; Soltis et al. 2004; Stefanović, Rice and Palmer 2004; Martin et al. 2005; Lockhart and Penny 2005). Alignment and orthology assignments are straightforward for the majority of coding regions, making plastid genomes ideal for phylogenetic reconstruction and studies of molecular evolution. Whole genome sequencing of plastid genomes provide copious data for testing hypothesized organismal relationships, comparing models of molecular evolution and developing analytical methodologies. At this point, however, with few plastid genomes available for analysis, care mast be taken to avoid being mislead by the results of some analyses. With nearly 40,000 sites in our ungapped nucleotide alignments of 61 genes, any method that is susceptible to statistical inconsistency may be affected by long-branch attraction.

Our results lead us to reject all but two hypotheses concerning the basal-most extant angiosperm lineages. The ML analyses provide weak support for *Amborella* and Nymphaeales as a clade sister to all other angiosperms, but the more popular hypothesis placing *Amborella* alone as sister to all other angiosperms could not be rejected. The monocots clearly constitute a slightly younger clade in the angiosperm phylogeny, although we estimate that the divergence of monocot and eudicot lineages occurred only 16 million years after the most recent common ancestor of extant angiosperms (table 4). We conclude that a lineage-specific increase in nucleotide substitution rates on the branch leading to the grasses and incomplete taxon sampling in the monocots confounded the analyses of Goremykin et al. (2003; 2004), resulting in the inference of the grasses sister to all other angiosperms in all analyses that did not include among-site rate variation. The earlier studies did place the branching point for *Amborella* or *Amborella* + Nymphaeales at the basal angiosperm node when ML analyses included a correction of among-site rate variation (Goremykin et al. 2003, 2004; Stefanović, Rice and Palmer 2004). With additional monocots included in the data matrix (table 1), we find that all MP and ML analyses of our 61-gene alignments (both nucleotides and amino acids) placed the branching point(s) for *Amborella* and the Nymphaeales at the base of the angiosperm phylogeny with strong support. Some of the distance-based NJ analyses placed core eudicots or eudicots as sister to the remaining angiosperms. However, all but the LogDet analyses of the amino acid alignment provided strong support for a *Amborella* + water lilies clade as sister to all other extant angiosperms when the analyses were restricted to seed plants (supplemental figures 1 and 4). The basis of differences in the results of LogDet analyses performed on the nucleotide and amino acid alignments deserves further investigation.

As has been described previously (e.g., Eyre-Walker and Gaut 1997), the substitution rate for chloroplast genes is accelerated both within the grasses and on the branch leading to the most common ancestor of maize, rice and wheat (fig. 3). The earlier work found an increase in synonymous substitution after the divergence of the grasses and the palms (Arecaceae, Arecales).

The phylograms shown in Figures 1, 3 and 4 suggest that the rate acceleration occurred within the Poales, after divergence from the most recent common ancestor of *Typha* (Typhaceae, Poales) and the grasses.

The phylogenetic position of the Nymphaeales relative to *Amborella* and the remaining angiosperms remains unresolved. While the parsimony analysis suggests strong support for *Amborella* as sister to all other angiosperms, the parametric bootstrap analyses performed here and in a previous study (Zanis et al. 2002) lead us to interpret the parsimony results cautiously. At the same time, the strong support for a *Amborella* + Nymphaeales clade observed in the NJ analysis is tempered by the moderate support for this clade in the ML analysis and the miniscule, nonsignificant difference in likelihood scores between the topologies A and B in figure 1. Whether the topology represented in figure 1A or 1B is correct, the molecular clock analysis suggests that the Nymphaeales lineage diverged from a sister lineage leading either to *Amborella* or to all other angiosperms some 25 million years after the most recent common ancestor of all extant angiosperms. To put this into context, the estimated dates for the youngest nodes on the ML phylogeny, *Nicotiana*/*Atropa* (10.56±0.03) and *Zea*/*Saccharum* (8.87 ± 0.01), are just under half this age.

Previous studies have reached different conclusions concerning the relationship of *Amborella* and the Nymphaeales at the base of the angiosperm phylogeny. The initial identification of *Amborella*, the Nymphaeales and the Austrobaileyales as the most basal lineages of extant angiosperms (Soltis, Soltis and Chase 1999; Qiu et al. 1999; Mathews and Donoghue 1999; Parkinson, Adams and Palmer 1999; Graham and Olmstead 2000) was a landmark event in molecular systematics. Although *Amborella* was favored as sister to all other extant angiosperm lineages in these seminal multi-gene studies (hypothesis ***a***, fig. 1A), a hypothesis placing *Amborella* with Nymphaeales in a clade sister to the remaining angiosperm lineages (hypothesis ***b***, fig. 1B) could not be rejected (Qiu et al. 2000; Mathews and Donoghue 2000; Parkinson, Adams and Palmer 1999). Barkman et al. (2000) favored the *Amborella* + Nymphaeales basal clade hypothesis after performing a series of MP, ML and NJ analyses on partitioned and complete multigene datasets, both before and after applying a controversial “noise-reduction” screen designed to identify and remove sites that may obscure phylogenetic signal (Lyons-Weiler, Hoelzer and Tausch 1996). However, some of their analyses provided strong support for hypothesis **a** and many gave weak to moderate support for either hypothesis. They used a nonparametric bootstrap resampling procedure test their prediction that as larger amounts of data were gathered from genes with similar evolutionary dynamics to those sampled, the support for the method-dependent conflict would grow increasingly strong due to statistical inconsistency.

Zanis et al. (2002) also found that inference of the relationship between *Amborella* and Nymphaeales was dependent on data partition and phylogenetic method. In general agreement with the previous studies of DNA sequence data, MP and nuclear ribosomal genes offered the strongest support for hypothesis **a**, while ML analyses of protein coding genes, and genes sampled from the chloroplast and mitochondrial genomes gave weak support for hypothesis ***b***. MP and ML analyses of the combined data set gave rather strong support for hypothesis a, but hypothesis ***b*** could not be rejected in a likelihood ratio test. This result is very similar to our finding of 63% bootstrap support for hypothesis ***b*** while hypothesis ***a*** cannot be rejected in the S-H test. Zanis et al. found as we did that parametric bootstrap analyses demonstrated bias in MP reconstruction toward recovery of *Amborella* as sister to all other extant angiosperms.

The total number of nucleotides included in this study was over 2.5 times the number analyzed by any of the previous multigene studies, yet we are still unable to conclusively reject either hypothesis ***a*** or ***b*** (fig. 1). Accurate phylogenetic resolution can generally be achieved more efficiently when taxa can be added to break long branches on the phylogeny (Graybeal 1998; Zwickl and Hillis 2002; Pollock et al. 2002; Hillis et al. 2003). This generalization, however, may not always apply to the resolution of branching order among basal lineages (Simmons and Miya 2004). Recent studies with extensive taxon sampling (Soltis, Soltis and Chase 1999; Zanis et al. 2002 [data set 2]; Hilu et al. 2003) have supported hypothesis ***a*** with moderate to high levels of support in MP analyses. It is not clear, however, whether increased taxon sampling was sufficient in these studies to interrupt the long branches responsible for the bias observed in the parametric bootstrap analyses performed here and by Zanis et al. (2002).

The results presented by Goremykin et al. (2003; 2004) demonstrate how incomplete taxon sampling can result in strong support for erroneous topological relationships. Graham and Olmstead (2000) had previously shown how sampling among basal angiosperm lineages can influence phylogenetic reconstruction of branching order. Their MP analysis found strong support (96%) for Nymphaeales as sister to all other angiosperms (e.g. fig. 1C) when the order was represented by *Cabomba* (Cabombaceae) alone, but when *Nymphaea* and *Cabomba* were included in the analysis *Amborella* moved to the base of the angiosperm phylogeny with moderate support (69%). When we added data from our six new plastid genome sequences one at a time to the 61 gene nucleotide matrices, we found that the addition of monocots *Typha* and *Yucca* changed the strongly supported position of the grasses in MP phylogenies (supplemental fig. 5). As was found by Stefanovic, Rice and, Palmer (2004), the grasses and Acorus were placed as successive sister lineages to the rest of the angiosperms in MP analyses when Acorus and the grasses were the only monocots included in the analysis (supplemental fig. 5F). As reported by Goremykin et al. (2003; 2004) we found that ML analyses including variation across sites placed lineages leading to *Amborella* and the water lilies at the base of the angiosperm phylogeny irrespective of taxon set (supplemental fig. 6).

Due to extinctions, taxa are not available to reduce the length of the critical branch separating the angiosperms and gymnosperms, nor the terminal branch leading to *Amborella*. It is possible that the combination of these long branches and the short internode subtending the branching point for the Nymphaeales and its sister lineage (*Amborella* [hypothesis ***b***] or the rest of the extant angiosperms [hypothesis ***a***]) may not allow us to conclusively reject hypothesis **a** or **b**. The possibility of long-branch attraction under these circumstances is expected to be especially problematic in analyses using only rapidly evolving coding (Hilu et al. 2003) and noncoding

(Borsch et al. 2003) sequences. The addition of species within the Cabombaceae (Nymphaeales), the Austrobaileyales, and magnoliids to the 61 gene dataset however, may lead to resolution of the relationships of *Amborella*, the Nymphaeales and the remaining extant angiosperms. Aside from the improvement in phylogenetic analyses based on nucleotide and amino acid substitutions, the addition of these and outgroup taxa to the 61 gene data matrix will likely improve alignment of gapped regions and increase the number of unambiguously scored indel characters.

Although all of the prior phylogenetic analyses of whole chloroplast genome sequence have focused on DNA or protein sequence analyses, Graham and Olmstead (2000) and Graham et al. (2000) showed that solid phylogenetic inference can be derived from careful characterization of insertion and deletion mutations in chloroplast genomes. Although much less numerous than base substitutions within most coding regions, our results agree with those of Graham et al, (2000) supporting evidence that these characters have much lower homoplasy than base substitutions and may provide a special collection of evidence bearing on branching events that are otherwise challenging to resolve due to phylogenetic artifacts such as long branch attraction (Rokas and Holland 2000; Graham et al. 2000). The unambiguous placement of *Amborella* and Nymphaeales as the most basal angiosperm lineages was resolved with our indel matrix even with the limited taxon sampling employed by Goremykin et al. (2003, 2004) (supplement fig. 7). The fact that the microstructural data does not appear to be affected by the same long-branch attraction problems is noteworthy, since it is clear (fig. 5B) that the rate of indel evolution is also dramatically increased in the lineage leading to the Poaceae. This may justify an extensive effort to identify more microstructural characters in these genomes.

We noted in our alignments many other regions that were rich in indel mutations, but where unambiguous character assignment was not yet possible in our judgment. It is likely that as additional genomes are sequenced, alignment of these difficult regions will improve, allowing the coding of many additional microstructural characters. Many of the lineage-specific indels that were ignored in this study should then emerge as synapomporphies among additional taxa.

These additions should also help resolve the other difficult nodes in the phylogenies involving the branching order of the magnoliids, monocots and eudicots. Relationships among these three clades, *Ceratophyllum* and Chloranthaceae have not been well resolved in previous studies, but the high support for the monophyly of the magnoliids (Magnoliales, Laurales, Canellales and Piperales) in the 17 gene analysis of Graham and Olmstead (2000) suggests that whole chloroplast genome sequences could provide enough phylogenetically informative nucleotide variation to clarify relationships among these taxa.

The eudicots (tricolpates) comprise roughly 64% of angiosperm species diversity (Judd and Olmstead 2004). While many nodes within the phylogeny for the group are well-supported, rapid diversification has made resolution of some nodes quite difficult. Resolution of the relationships among the major core eudicot lineages, including the Caryophyllales, rosids and asterids, has been particularly recalcitrant. The moderate to high bootstrap support observed for a clade joining the spinach lineage (Amaranthaceae, Caryophyllales) with *Atropa* and *Nicotiana* (Solanaceae, Solanales, euasterid I), should be interpreted cautiously. The chloroplast genome of *Panax* (Araliaceae, Apiales, euasterid II) has recently been published (Kim and Lee 2004) and its inclusion in the 61 gene data set results in slightly reduced bootstrap support for a Caryophyllales + asterid clade (supplemental fig. 8).

Over the last decade, advances in our understanding of phylogenetic relationships among extant angiosperms (Soltis and Soltis 2004b; Judd and Olmstead 2004; Chase 2004) have provided an improved framework for comparative analyses designed to elucidate the evolution of important features ranging from endosperm development (Williams and Friedman 2002) to MADS box gene evolution (Becker and Theissen 2003; Litt and Irish 2003; Stellari Jaramillo and Kramer 2004; Kramer Jaramillo and DiStillo 2004; Kim et al. 2004) to the evolution of floral perianth organization (Zanis et al. 2003; Soltis et al. 2004b). Just as favored evolutionary scenarios had to be abandoned with the demise of the anthophyte hypothesis (Goremykin et al. 1996), inferences drawn in these and many other comparative studies would have had to be reexamined if the position of the monocots recovered in the phylogenies of Goremykin et al. (2003; 2004) were supported in subsequent studies. This study and others (Soltis and Soltis 2004a; Stefanović, Rice and Plamer 2004), however, have tested and rejected the hypothesized position of monocots sister to all other angiosperm lineages (Goremykin et al. 2003; 2004).

As genome scale sequencing and phylogenetic analyses become more common, the possible influence of long branch attraction must be seriously considered in any interpretation of the resulting phylogenies. Phylogenies based on many genes sampled from a few model species will be especially susceptible to long-branch attraction (Soltis et al. 2004a; Philippe, Lartillot and Brinkmann 2005). We contend that genome-scale phylogenetic studies can avoid misinterpretation of artifactual results by employing parametric bootstrap analyses (e.g., Sanderson et al. 2000; Zanis et al. 2002), multiple reconstruction methods (MP, ML, NJ, Bayesian), a variety of models of molecular evolution (e.g. Stefanović, Rice and Palmer 2004; Phillips et al 2004), consideration of variation in substitution patterns among lineages (Lockhart et al. 1998; 1999; Lopez, Casane and Philippe 2002; Huelsenbeck 2002; Phillips, Delsuc, and Penny 2004; Ané et al. 2005; Kolaczkowski and Thornton 2004; Spencer Susko and Roger 2005; Martin et al. 2005), taxon subsampling (Graham and Olmstead 2000; Soltis and Soltis 2004a) and analyses of multiple data partitions (e.g., Barkman et al. 2000; Zanis et al. 2002). Inconsistencies among results derived from different approaches should be examined and explained rather than ignored.

Acknowledgments

We wish to thank Sean Graham, Jeff Palmer, Bill Martin, David Penny, Doug Soltis and an anonymous reviewer for helpful discussions concerning this work. Funding was provided by the National Science Foundation grants DEB-**0120709** and DBI-**0115684.** Part of this work was performed under the auspices of the U.S. Department of Energy, Office of Biological and

Environmental Research, by the University of California, Lawrence Berkeley National Laboratory, under contract No. DE-AC03-76SF00098.

References

Ané, C., J. G. Burleigh, M. M. McMahon, and M. J. Sanderson. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol **22**:914-924.

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

Barkman, T. J., G. Chenery, J. R. McNeal, J. Lyons-Weiler, W. J. Ellisens, G. Moore, A. D. Wolfe, and C. W. dePamphilis. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc Natl Acad Sci U S A **97**:13166-13171.

Becker, A., and G. Theissen. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phylogenet Evol **29**:464-489.

Borsch, T., K. W. Hilu, D. Quandt, V. Wilde, C. Neinhuis, and W. Barthlott. 2003. Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol **16**:558-576.

Chase, M. W. 2004. Monocot relationships: an overview. Am. J. Bot. **91**:1645-1655.

Chaw, S. M., C. C. Chang, H. L. Chen, and W. H. Li. 2004. Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes. J Mol Evol **58**:424-441.

Crane, P. R., E. M. Friis, and K. R. Pederson. 1995. The origin and early diversification of angiosperms. Nature **374**:27-33.

Davies, T. J., T. G. Barraclough, M. W. Chase, P. S. Soltis, D. E. Soltis, and V. Savolainen. 2004. Darwin's abominable mystery: Insights from a supertree of the angiosperms. Proc Natl Acad Sci U S A **101**:1904-1909.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research **8**:186-194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.

Eyre-Walker, A., and B. S. Gaut. 1997. Correlated rates of synonymous site evolution across plant genomes. Mol Biol Evol **14**:455-460.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Systematic Zoology **27**:401-410.

Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-791.

Felsenstein, J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.*

Gordon, D., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res 8:195-202.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol **21**:1445-1454.

Goremykin, V. V., K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that amborella is not a basal angiosperm. Mol Biol Evol **20**:1499-1505.

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. Mol Biol Evol. **13**:383-396.

Graham, S. W., and R. G. Olmstead. 2000. Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am. J. Bot. **87**:1712-1730.

Graham S.W., P.A. Reeves, A.C.E. Burns and R.G. Olmstead. 2000. Microstructural changes in noncoding chloroplast DNA: Interpretation, evolution, and utility of indels and inversions in basal angiosperm phylogenetic inference. International Journal Of Plant Sciences **161**:S83-S96

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends Genet **20**:80-86.

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

Guindon, S., and O. Gascuel. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52**:696-704.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol **22**:160-174.

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

Hillis, D. M., D. D. Pollock, J. A. McGuire, and D. J. Zwickl. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst Biol **52**:124-126.

Hilu, K. W., T. Borsch, K. Muller, D. E. Soltis, P. S. Soltis, V. Savolainen, M. W. Chase, M. P. Powell, L. A. Alice, R. Evans, H. Sauquet, C. Neinhuis, T. A. B. Slotta, J. G. Rohwer, C. S. Campbell, and L. W. Chatrou. 2003. Angiosperm phylogeny based on matK sequence information. Am. J. Bot. **90**:1758-1776.

Hiratsuka, J., H. Shimada, R. Whittier, T. Ishibashi, M. Sakamoto, M. Mori, C. Kondo, Y. Honji, C. R. Sun, B. Y. Meng, and et al. 1989. The complete sequence of the rice (Oryza sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol Gen Genet **217**:185-194.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol **19**:698-707.

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

Jansen, R.K., L.A. Raubeson, J.L. Boore, C.W. dePamphilis, T. Chumley, R.C. Haberle, S.K. Wyman, A.J. Alverson, R. Peery, S.J. Herman, H.M. Fourcade, J.V. Kuehl, J.R. McNeal, J.H. Leebens-Mack, and L. Cui. 2005. Methods for Obtaining and Analyzing Whole Chloroplast Genome Sequences. Methods in Enzymology, *In press*.

Jones, D. T., W. R. Taylor, and J. M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci **8**:275-282.

Judd, W. S., and R. G. Olmstead. 2004. A survey of tricolpate (eudicot) phylogenetic relationships. Am. J. Bot. **91**:1627-1644.

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

Kellogg, E. 2000. Evolutionary History of the Grasses. Plant Physiol. **125**: 1198-1205

Kim,K.-J. and Lee,H.-L. 2004. Complete chloroplast genome sequence from korea ginseng (*Panax schinseng* Nees) and comparative analysis of sequence evolution among 17 vascular plants. DNA Res. **11**:247-261.

Kim, S., M.-J. Yoo, V. A. Albert, J. S. Farris, P. S. Soltis, and D. E. Soltis. 2004. Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. Am. J. Bot. **91**:2102-2118.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature **431**:980-984.

Kramer, E. M., M. A. Jaramillo, and V. S. Di Stilio. 2004. Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. Genetics **166**:1011-1023.

Kugita, M., Y. Yamamoto, T. Fujikawa, T. Matsumoto, and K. Yoshinaga. 2003a. RNA editing in hornwort chloroplasts makes more than half the genes functional. Nucleic Acids Res **31**:2417-2423.

Kugita, M., A. Kaneko, Y. Yamamoto, Y. Takeya, T. Matsumoto, and K. Yoshinaga. 2003b. The complete nucleotide sequence of the hornwort (*Anthoceros formosae*) chloroplast genome: insight into the earliest land plants. Nucleic Acids Res **31**:716-721.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.

Litt, A., and V. F. Irish. 2003. Duplication and diversification in the APETALA1/FRUITFULL floral homeotic gene lineage: implications for the evolution of floral development. Genetics **165**:821-833.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realisticmodel of sequence evolution. Mol. Biol. Evol. **11**:605–612.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. H. Huson, M. A. Charleston, and C. J. Howe. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol **15**:1183-1188.

Lockhart, P. J., C. J. Howe, A. C. Barbrook, A. W. D. Larkum, and D. Penny. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol **16**:573-576.

Lockhart, P. J., and D. Penny. 2005. The place of Amborella within the radiation of angiosperms. Trends Plant Sci **10**:201-202.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol Biol Evol **19**:1-7.

Lyons-Weiler, J., G.A. Hoelzer and R.J. Tausch. 1996. Relative apparent synapomorphy analysis (RASA). I: the statistical measurement of phylogenetic signal. Mol. Biol. Evol. **13**:749–757.

Maier, R.M., K. Neckermann, G.L. Igloi, and H. Kossel. 1995. Complete sequence of the maize chloroplast genome: gene content, hotspots of divergence and fine tuning of genetic information by transcript editing J. Mol. Biol. **251**:614-628.

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

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

Mathews S. and M. J. Donoghue 2000 Basal angiosperm phylogeny inferred from duplicate phytochromes A and C. International Journal of Plant Sciences **161**:S41-S41

Ohyama, K., H. Fukuzawa, T. Kohchi, T. Sano, S. Sano, H. Shirai, K. Umesono, Y. Shiki, M. Takeuchi, Z. Chang, and et al. 1988. Structure and organization of Marchantia polymorpha chloroplast genome. I. Cloning and gene identification. J Mol Biol **203**:281-298.

Parkinson, C. L., K. L. Adams, and J. D. Palmer. 1999. Multigene analyses identify the three earliest lineages of extant flowering plants. Curr Biol **9**:1485-1488.

Philippe, H., N. Lartillot, and H. Brinkmann. 2005. Multigene Analyses of Bilaterian Animals Corroborate the Monophyly of Ecdysozoa, Lophotrochozoa and Protostomia. Mol Biol Evol. in press.

Philippe H., and J. Laurent. 1998. How good are deep phylogenetic trees? Curr. Opin. Genet. Dev. **8**: 616-623.

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

Pollock, D. D., D. J. Zwickl, J. A. McGuire, and D. M. Hillis. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst Biol **51**:664-671.

Qiu, Y. L., J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. 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., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci **13**:235-238.

Rokas, A., and P.W. Holland. Rare genomic changes as a tool for phylogenetics. Trends Ecol Evol. **15**:454-459

Sanderson, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**:301-302.

Sanderson, M. J., J. L. Thorne, N. Wikstrom, and K. Bremer. 2004. Molecular evidence on plant divergence times. Am. J. Bot. **91**:1656-1665.

Sanderson, M. J., M. F. Wojciechowski, J. M. Hu, T. S. Khan, and S. G. Brady. 2000. Error, bias, and long-branch attraction in data for two chloroplast photosystem genes in seed plants. Mol Biol Evol **17**:782-797.

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

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

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

Schneider, H., E. Schuettpelz, K. M. Pryer, R. Cranfill, S. Magallon, and R. Lupia. 2004. Ferns diversified in the shadow of angiosperms. Nature **428**:553-557.

Shimodaira, H. and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114-1116.

Shinozaki,K., M. Ohme, M Tanaka, T. Wakasugi, N. Hayashida, T. Matsubayashi, N. Zaita, J. Chunwongse, J. Obokata, K. Yamaguchi-Shinozaki, C. Ohto, K. Torazawa, B.Y. Meng, M. Sugita, H. Deno, T. Kamogashira, K. Yamada, J. Kusuda, F. Takaiwa, A. Kato, N. Tohdoh, H. Shimada, and M. Sugiura. 1986. The complete nucleotide sequence of tobacco chloroplast genome: its gene organization and expression EMBO J. **5**:2043-2049.

Simmons M.P., and M. Miya. 2004. Efficiently resolving the basal clades of a phylogenetic tree using Bayesian and parsimony approaches: a case study using mitogenomic data from 100 higher teleost fishes. Mol Phylogenet Evol. **31**:351-362

Soltis, D. E., V. A. Albert, V. Savolainen, K. Hilu, Y. L. Qiu, M. W. Chase, J. S. Farris, S. Stefanović, D. W. Rice, J. D. Palmer, and P. S. Soltis. 2004a. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci **9**:477-483.

Soltis, D. E., V. A. Albert, S. Kim, M-J. Yoo, P. S. Soltis, P. S., M. W. Frohlich, J. H. Leebens-Mack, H., Kong, P. K., Wall, H., Ma, and C. W. dePamphilis, 2004b Evolution of the Flower. *in* R. Henry, ed. Diversity and Evolution of Plants. CABI Publishers

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

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

Soltis, P. S., D. E. Soltis, and M. W. Chase. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature **402**:402-404.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, Parsimony, and Heterogeneous Evolution. Mol Biol Evol **22**:1161-1164.

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

Steel, M. A. 1994. Recovering a tree from the leaf colorations it generates under a Markov model. Appl. Math. Lett. **7**:19–23.

Stellari, G. M., M. A. Jaramillo, and E. M. Kramer. 2004. Evolution of the APETALA3 and PISTILLATA lineages of MADS-box-containing genes in the basal angiosperms. Mol Biol Evol **21**:506-519.

Sugiura, C., Y. Kobayashi, S. Aoki, C. Sugita, and M. Sugita. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of rpoA from the chloroplast to the nucleus. Nucleic Acids Res **31**:5324-5331.

Swofford, D. L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Thollesson, M. 2004. LDDist: a Perl module for calculating LogDet pair-wise distances for protein and nucleotide sequences. Bioinformatics **20**:416-418.

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 Res **22**:4673-4680.

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

Williams, J. H., and W. E. Friedman. 2002. Identification of diploid endosperm in an early angiosperm lineage. Nature **415**:522-526.

Wolf, P. G., C. A. Rowe, R. B. Sinclair, and M. Hasebe. 2003. Complete nucleotide sequence of the chloroplast genome from a leptosporangiate fern, *Adiantum capillus-veneris* L. DNA Res **10**:59-65.

Wyman, S. K., R. K. Jansen, and J. L. Boore. 2004. Automatic annotation of organellar genomes with DOGMA. Bioinformatics **20**:3252-3255.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol **39**:306-314.

Zanis, M. J., D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. 2002. The root of the angiosperms revisited. Proc Natl Acad Sci U S A **99**:6848-6853.

Zanis, M. J., P. S. Soltis, Y.-L. Qiu, E. Zimmer, and D. E. Soltis. 2003. Phylogenetic analyses and perianth evolution in basal angiosperms. Annals of the Missouri Botanical Garden **90**:129-129.

Zwickl, D. J., and D. M. Hillis. 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst Biol **51**:588-598.

Table 1. GenBank accession numbers for the sequences included in this study

Taxon (Collection local; voucher ID)	GenBank Accession Numbers	Reference
Bryophytes		
Anthoceros formosae	NC_004543	Kugita et al. 2003
Physcomitrella patens	NC_005087	Sugiura et al. 2003
Marchantia polymorpha	NC_001319	Ohyama et al. 1986
Ferns and allies		
Psilotum nudum	NC_003386	Wakasugi et al., unpublished
Adiantum capillus-veneris	NC_004766	Wolf et al. 2003
Gymnosperms		
Pinus thunbergii	NC_001631	Wakasugi et al. 1994
Ginkgo biloba (Travis Co., TX; RCH155 TEX)	xxx-xxx	Current study
Basal-most angiosperm lineages[a]		
Amborella trichopoda	NC_005086	Goremykin et al. 2003
Nymphaea alba	NC_006050	Goremykin et al. 2004
Nuphar advena (Centre Co., PA; PAC)	xxx-xxx	Current study
Monocots		
Acorus americanus (Crawford Co., PA; jlm-acor001 PAC)	xxx-xxx	Current study
Typha latifolia (Yavapai CO., AZ; RCH188 TEX)	xxx-xxx	Current study

Yucca schidigera (San Diego Co., CA; jlm-yuc370 PAC)	xxx-xxx	Current study
Saccharum officinarum	NC_006084	Asano et al. 2004
Zea mays	NC_001666	Maier et al. 1995
Oryza sativa	NC_001320	Hiratsuka et al. 1989
Triticum aestivum	NC_002762	Ikeo and Ogihara, unpublished
Magnoliids		
Calycanthus floridus	NC_004993	Goremykin et al. 2003
Eudicots		
Ranunculus macranthus (Travis Co., TX; RCH1184 TEX)	xxx-xxx	Current study
Nicotiana tabacum	NC_001879	Shinozaki et al. 1986
Atropa belladonna	NC_004561	Schmitz-Linneweber et al. 2002
Spinacia oleracea	NC_002202	Schmitz-Linneweber et al. 2001
Lotus corniculatus	NC_002694	Kato et al. 2000
Medicago truncatula	NC_003119	Lin et al., unpublished
Arabidopsis thalliana	NC_000932	Sato et al. 1999
Oenothera elata	NC_002693	Hupfer et al. 2000

[a] basal angiosperm lineages as determined in most molecular systematic studies since 1999 (see text).

Table 2. The lineage(s) that is inferred to be most basal in angiosperms was ambiguous in most single gene analyses. The identities of basal lineages inferred with at least 50% bootstrap support in the MP analyses are shown for the nucleotide and amino acid alignments with all gapped positions removed. Bootstrap support (%) for basal position indicated lineage is shown in square brackets for each gene.

Taxon[1]	Codon Positions 1+2+3	Codon Positions 1+2	Amino Acids
Poaceae+*Oenothera*	1 (clpP [80])		
Poaceae		3 (rpoB [50], psbK [54], rps3 [53])	2 (rpoC1 [86] rps12 [64])
Typha			1 (psbB [53])
Poales (*Typha*+Poaceae)		1 (psbB [53])	
Eudicots	1 (rps2 [82])		
Core Eudicots	2 (psbB [53], psbD [54])		
Oenothera		1 (rps2 [50])	1 (rps2 [69])
Spinacia		1 (petD [53])	1 (rps15 [66])
Amborella	5 (atpE [61], atpF [56], psaA [59], rbcL [59], rpoA [86])	3 (atpE [58], rpoA [64], cemA [68])	2 (atpE [57], cemA [64])

Nymphaeales	1 (rpoB [69])		
Nymphaeales + *Amborella*	3 (rpoC1 [66], rpoC2 [74], rps4 [65])	2 (rpoC2 [73], matK [51])	1 (ccsA [53])
Unresolved	48	50	54
Total	61	61	61

Table 3. Parametric bootstrap results show that parsimony analyses are more subject to long branch attraction than the model based likelihood and neighbor joining analyses when datasets are simulated on phylogenies B (Amborella +Nymphaeales basal clade) and C (Nymphaeales basal-most clade). Parsimony and neighbor joining performed poorly when data were simulated on the ML tree forcing monocots as sister to all other angiosperms. Rate of recovering the simulated topology is shown in bold for each reconstruction method and simulated phylogeny.

Simulated phylogeny[a]	Likelihood				Parsimony				Neighbor Joining			
	Amb	Amb +Nym	Nym	Monocot	Amb	Amb +Nym	Nym	Monocot	Amb	Amb +Nym	Nym	Monocot
A	**100**	--	--		**100**	--	--		**100**	--	--	
B	--	**100**	--		21	**79**	--		-	**100**	--	
C	--	--	**100**		10	--	**90**		1	1	**98**	
D[b]		1		**98**		48		**4**		29		**41**

[a] Input phylogenies for the simulations are shown in Figure 1.

[b] Rows do not sum to one for tree D because other angiosperm rootings were observed in the in the bootstrap trees, including monocots+*Amborella*+Nymphaeales, *Calycanthus*+monocots+*Amborella*+Nymphaeales and *Calycanthus* as sister to remaining angiosperms.

Table 4. Estimated ages (in millions of years) for nodes labeled in figure 6.-need to update. The median estimates are shown with the range of observed estimates for 200 bootstrap replicates. All estimates were within one million years for the nodes with no range shown.

Node Label	Description	Constraint	Estimated Age
a	euphyllophytes	380 minimum 401 maximum	410
b	seed plants		334
c	*Pinus* + *Ginkgo*	310 minimum	310
d	angiosperms	132 minimum	161 (158-165)
e	Nymphaeales/*Amborella*		136 (134-165)
f	Nymphaeaceae s.s.		22 (21-23)
g	magnoliids/monocots/eudicots		145 (143-147)
h	monocots		133 (131-135)
i	Asparagales/commelinids		117 (116-118)
j	Poales		107 (106-109)
k	core Poaceae	55 minimum	55
L	eudicots	125 fixed	125
m	core eudicots		113
n	rosids		108 (108-109)
o	Faboideae		61 (61-62)
p	asterids/Caryophyllales		105
q	Solanaceae		12

Figure Legends:

Figure 1. Nucleotide-based ML phylogenies estimated using the HKY+Γ+I substitution model while constraining *Amborella* (A), *Amborella*+water lilies (B), water lilies (C) or monocots (D) as sister to remaining angiosperms. Likelihood for each hypothesis is shown with each phylogeny.

Figure 2. Pair-wise HKY distances estimated for third codon positions are linearly correlated with distances estimated for first and second positions.

Figure 3. Comparison of unrooted MP phylogenies estimated from amino acid alignments with the taxon set analyzed by Goremykin et al. (2004; A) and in this study (B) reveal the long branch leading to grasses.

Figure 4. Single MP (A), ML (B), and NJ (C) phylogenies estimated from the 61-gene nucleotide alignment are shown. Most nodes on each phylogeny were recovered in 100% of the bootstrap replicates and only values < 100% are shown for each node. All analyses place *Amborella* and the water lilies as basal lineages in the angiosperm phylogeny.

Figure 5. Parsimony analysis of microstructural mutations in 61 coding regions support *Amborella* and water lilies as the most basal lineages of angiosperms: A) bootstrap consensus of 116 parsimony-informative insertion and deletion characters, B) phylogram of one of 12 most parsimonious trees (141 steps) with branch lengths drawn proportional to the number of inferred insertion and deletion mutations on each branch.

Figure 6. Six phylogenetically informative insertion-deletion mutations bearing on the position of *Amborella* and water lilies relative to other angiosperms. Each mini-alignment is identified by the gene name and position in the *Amborella* sequence of the first amino acid in the indel of interest. Four indels (A, B, C, and D) support the basal position of *Amborella* and water lilies;

one indel (E) supports *Amborella* as the sole basal-most angiosperm lineage; one indel supports the monophyly of *Amborella* plus water lilies. Thus, characters E and F are consistent with A-D, but conflict with regard to the relationship between *Amborella* and water lilies. No phylogenetically informative indel characters were observed that would have supported a basal position for grasses or monocots in the angiosperms.

Figure 7. The ML phylogeny inferred from the nucleotide analysis with labeled nodes corresponding to the minimum divergence time estimates shown in table 4.

Supplemental Figure 1. ML and NJ phylogenies recovered using the HKY substitution model without correction for among-site variation. Comparison of ML (A and B) and NJ phylogenies (C and D) estimated from alignments including euphyllophytes (A and C) or just seed plant sequences (B and D) shows that the NJ analyses run under the simple HKY model are influenced by the inclusion or exclusion of distant outgroup sequencess

Supplemental Figure 2. ML (A), MP (B) and NJ (C) phylogenies with bootstrap values from analyses of 1st and 2nd codon positions in the nucleotide alignments are very similar to those estimated with all three codon positions (fig. 4). The bootstrap value increased for the water lilies + *Amborella* clade in the ML analysis and the poorly supported placement of *Calycanthus* relative to the eidicots and monocots changed in the MP analysis. ML and NJ analyses performed using the HKY+Γ+I model as described in text. Bootstrap values are not shown for branches with 100% support.

Supplemental Figure 3. The results of ML (A) and NJ (C) analyses of amino acid alignment differ slightly from those estimated with the complete nucleotide alignment (fig. 4). The ML analysis (JTT+Γ+I) returns poor support *Amborella* as sister to all other angiosperms, and the NJ analysis (JTT+Γ) places eudicots as sister to the remaining angiosperms. A NJ analysis restricted to the seed plants returns relationships identical to those for seed plants shown in figure 4C.

Supplemental Figure 4. Phylogenies resulting from analyses performed on the complete nucleotide and amino acid alignments (ungapped) using LogDet corrected distances place to core eudicots as sister all other angiosperms including *Ranunculus* (A-D). Whereas analyses restricted to the seed plant nucleotide alignment recover relationships identical to those for seed plants shown in figure 4C (E and F), the eudicots are placed sister to the remaining angiosperms in analyses of the amino acid alignment for seed plants (G and H). Topologies are identical for LogDet analyses performed with (B, D, F and H) and without (A, C, E and G) variation rates across sites (see text).

Supplemental Figure 5. Phylogenies from MP analyses adding *Ginkgo* (A), *Nuphar* (B), *Ranunculus* (C), *Acorus* (D), *Yucca* (E) and *Typha* (F) one at a time to a 61 gene nucleotide alignment of previously available plastid genomes.

Supplemental Figure 7. Phylogenies from ML analyses adding *Ginkgo* (A), *Nuphar* (B), *Ranunculus* (C), *Acorus* (D), *Yucca* (E) and *Typha* (F) one at a time to a 61 gene nucleotide alignment of previously available plastid genomes. The HKY+Γ+I substitution model was used in all analyses.

Supplemental Figure 7. Parsimony bootstrap consensus trees of indel characters, using taxon sets from A) Goremykin et al. (2003) and B) Goremykin (2004). For analysis A, one MP tree was obtained (115 steps; CI=0.9111; RC = 0.8856) with *Amborella* the first branching angiosperm, while B obtained two MP trees at 127 steps (CI=9134; RC = 0.8560), one with *Amborella* and one with *Amborella + Nymphaea* as the earliest angiosperm branch.

Supplemental Figure 8. Bootstrap consensus phylogenies for ML, MP and NJ analyses of nucleotide alignment including 61 genes from *Panax schinseng* plastid genome sequence (Kim and Lee 2004) are consisten with those shown in figure 4. All analyses performed as described in text.

Supplemental Data Matrix 1. Nexus file with nucleotide alignment - http://chloroplast.cbio.psu.edu/

Supplemental Data Matrix 2. Nexus file with amino acid alignment - http://chloroplast.cbio.psu.edu/

Supplmental Data matrix 3. Nexus file with microstructural characters. http://chloroplast.cbio.psu.edu/