Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny

Todd J. Barkman[†‡§], Gordon Chenery[¶], Joel R. McNeal[†], James Lyons-Weiler[†], Wayne J. Ellisens[‖], Gerry Moore[††], Andrea D. Wolfe[‡‡], and Claude W. dePamphilis[†§]

[†]Department of Biology, Institute of Molecular Evolutionary Genetics, and Life Sciences Consortium, Pennsylvania State University, University Park, PA 16802; [¶]Montgomery Bell Academy, Nashville, TN 37205; [‖]Department of Botany and Microbiology, University of Oklahoma, Norman OK 73019; [††]Brooklyn Botanic Garden, Brooklyn, NY 11225; and [‡‡]Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, OH 43210

Communicated by Masatoshi Nei, Pennsylvania State University, University Park, PA, September 6, 2000 (received for review June 9, 2000)

Plant phylogenetic estimates are most likely to be reliable when congruent evidence is obtained independently from the mitochondrial, plastid, and nuclear genomes with all methods of analysis. Here, results are presented from separate and combined genomic analyses of new and previously published data, including six and nine genes (8,911 bp and 12,010 bp, respectively) for different subsets of taxa that suggest *Amborella* + Nymphaeales (water lilies) are the first-branching angiosperm lineage. Before and after tree-independent noise reduction, most individual genomic compartments and methods of analysis estimated the *Amborella* + Nymphaeales basal topology with high support. Previous phylogenetic estimates placing *Amborella* alone as the first extant angiosperm branch may have been misled because of a series of specific problems with paralogy, suboptimal outgroups, long-branch taxa, and method dependence. Ancestral character state reconstructions differ between the two topologies and affect inferences about the features of early angiosperms.

The origin of flowering plants and characteristics of angiosperm ancestors have long been pondered with little consensus having been reached to date (1, 2). The anthophyte hypothesis (ref. 3 and references therein), in which the closest living relatives of angiosperms are believed to be the Gnetales, a small and enigmatic group of gymnosperms, suggested a common origin for several shared characteristics such as double fertilization (4). However, recent molecular studies have opposed the anthophyte hypothesis, leaving flowering plants without a close extant relative and questioning the interpretation of Gnetales–angiosperm similarities (5–8). Fossil studies have been critical for inferring morphological characteristics of early angiosperms as well as their time of origin (9), but a more complete understanding awaits the discovery of additional material (10). Thus, extant angiosperm phylogeny provides critical evidence for improving inferences of early flowering plant characteristics. Recent reports that *Amborella* is the first-branching extant flowering plant have profoundly influenced our view of angiosperm relationships (11–14). In these reports, the second- and third-deepest branches of the estimated tree included water lilies (Nymphaeales) and a lineage of four families composed of the star-anise family (Illiciaceae), Schisandraceae, Austrobaileyaceae, and Trimeniaceae. These basal angiosperm relationships, based on evidence from the plastid (pt), mitochondrial (mt), and nuclear (nuc) genomes, represent evolutionary hypotheses that could profoundly enhance our understanding of the ancestral characters of angiosperms, but require corroboration.

The most complete approach for inferring plant phylogeny utilizes sequences residing in each genomic compartment. Although rates of nucleotide substitution (15), levels of recombination (16), and patterns of inheritance (17) differ among the mitochondrion, plastid, and nucleus, their genomes have coexisted within plant cells since before the origin of land plants (18) and are expected to trace the same evolutionary history. Therefore, well-supported congruent (19, 20) phylogenetic estimates from all three genomic compartments would result in the highest confidence of angiosperm relationships. Population- and organismal-level processes, such as lineage sorting (21) and horizontal gene transfer (22), can significantly alter the individual evolutionary histories of each genome; however, recent studies suggest that these processes may not be problematic for reconstructing deep angiosperm phylogenetic divergences (refs. 23 and 24; but see ref. 25 for an exception). Molecular-level processes such as gene duplication can also complicate inferences of phylogeny because analyses of mixed paralogs may result in inaccurate species tree estimation (26).

Not only is evolutionary history predicted to be identical for the three plant genomes, all methods of analysis are expected to converge on similar phylogenetic estimates (27). This expectation is based on results of simulation and experimental studies that have shown most methods tend to estimate phylogeny accurately (28–31). This is true except when the number of sites sampled is small (32) or the dataset causes inconsistency (33–36). Therefore, although topological estimates should agree, if estimates are misled (37, 38) an apparent pattern of incongruence between genomes or methods could result. Likewise, although the bootstrap (39) is generally interpreted as a measure of confidence in branching relationships (40), when methods are misled, significant support for incorrect nodes may result (41).

A potentially important step in phylogenetic analyses involves data exploration aimed at identifying problematic taxa or characters before tree estimation. Tests of normality are widely used before parameter estimation and hypothesis testing; however, only recently have analogous tools become available to phylogeneticists. Relative Apparent Synapomorphy Analysis (RASA; ref. 42) is a tree-independent method of data exploration for measuring phylogenetic signal that can be used to objectively choose optimal outgroups (43), identify and remove long-branch taxa (44), and even detect and reduce noise from a data set

Abbreviations: pt, plastid; mt, mitochondrial; nuc, nuclear; BP, bootstrap proportion; NJ, neighbor joining; ML, maximum likelihood; RASA, Relative Apparent Synapomorphy Analysis; tRASA, test statistic for phylogenetic signal; UW, unweighted parsimony; T, transversion parsimony; NJ-P, neighbor joining with *p* distance; HKY, Hasegawa–Kishino–Yano; GTR, general-time-reversible.

See commentary on page 12939.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AY009406–AY009456).

[‡]Present address: Department of Biological Sciences, Western Michigan University, Kalamazoo, MI 49008.

[§]To whom reprint requests may be addressed. E-mail: cwd3@psu.edu or todd.barkman@wmich.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Article published online before print: *Proc. Natl. Acad. Sci. USA*, 10.1073/pnas.220427497. Article and publication date are at www.pnas.org/cgi/doi/10.1073/pnas.220427497

Table 1. Datasets analyzed in this paper

Dataset	Genomic compartment			No. of sites, bp				No. of taxa
				Total		Shared variable		
	pt	nuc	mt	Raw	nr	Raw	nr	
Six genes	*rbcL*/*atpB*	18S	*matR*/*coxI*/*atpA*	8,911	8,087	1,374	878	35
Mathews and Donoghue (11)	—	*PHYA*/*PHYC*	—	1,104		634		26
Qiu *et al.* (12)	*rbcL*/*atpB*	18S	*matR*/*atpA*	8,733		2,330		105
Parkinson *et al.* (14)	*rbcL*	18S	*coxI*/19S/*rps2*	6,562		1,185		51
Nine genes	*rbcL*/*atpB*	18S/*PHYC*	*matR*/*coxI*/*atpA*/19S/*rps2*	12,112	12,010	1,316	931	15

Specific genes, genomic compartment, number of total and shared variable sites given in bp for both raw and noise-reduced (nr) data, and number of taxa are listed for each dataset.

before phylogenetic analysis. This process of data exploration, followed by optimal adjustments of taxon and character sampling, was shown to increase the probability of obtaining accurate phylogenetic estimates and has recently been used in empirical studies (7, 45–48). In this paper we present previously unpublished data, RASA analyses, and phylogenetic tree estimates that reveal striking discordance for basal angiosperm relationships depending on the method of analysis used.

Methods

Fifty mt DNA sequences of *atpA* (1,239 bp) and *coxI* (1,415 bp) were generated for this study by standard PCR methods followed (in most cases) by automated DNA sequencing on a Beckman-Coulter CEQ2000 genetic analyzer. Detailed protocols, GenBank accession numbers, voucher data, and detailed results of phylogenetic analyses (tree lengths, optimality scores, etc.) are available at http://depcla4.bio.psu.edu/basals. In cases where multiple copies of an amplified gene were detected, TA cloning (Invitrogen) was performed according to the supplier's specifications. All other sequences were obtained from previously published studies (11, 12, 14) and GenBank.

Table 1 lists the five datasets analyzed in this study. The six- and nine-gene datasets (which include our data) represent the largest datasets yet accumulated to study basal angiosperm phylogeny (Table 1). The primary analyses performed in this paper were focused on the six-gene dataset. This dataset included 33 putatively basal angiosperm species that were sequenced for all six genes (pt *rbcL* and *atpB*, nuc 18S rRNA, and mt *matR*, *atpA*, and *coxI*). Eight potential gymnosperm outgroups were considered even though *atpB* or *matR* was not sequenced for some of these species. A secondary analysis of nine genes for 15 taxa was performed with pt *rbcL* and *atpB*, nuc 18S, and phytochrome C (*PHYC*) and mt *matR*, *coxI*, *atpA*, *rps2*, and 19S rRNA (Table 1). Reanalyses of three published original data sets assembled for basal angiosperms (11, 12, 14) were performed to determine whether previous conclusions were method dependent and whether the estimates obtained in this study were taxon- or gene-sampling dependent.

Before tree estimation, RASA 2.4 (http://bio.uml.edu/LW/RASA.html) was used to measure phylogenetic signal (tRASA) with all gaps coded as "?". After phylogenetic signal was measured, taxon-variance ratios (44) were examined to screen for potential long-branch taxa. Optimal outgroup analysis (43) was performed for multiple outgroup assemblages. Once optimal outgroups were determined, noise reduction was performed. "Noise" is defined as any site that suppresses a measure of phylogenetic signal or hierarchy (tRASA in this case) of a dataset. Noisy sites, or discordant site configurations, could be present in a data set because of experimental errors (sequencing and alignment errors) and/or because of underlying processes that generated the data causing random error (e.g., saturation and recombination) and systematic error (e.g., selection and long-branch attraction). Noise reduction identifies a "noisy site" by examining its effect on the measure of phylogenetic signal of the entire data set. Specifically, when a noisy site is removed from a dataset, the overall measure of signal should increase because its presence conflicts with the predominant hierarchy (branching patterns) present in the dataset. The noise reduction routine in RASA 2.4 involves: (*i*) the calculation of tRASA (phylogenetic signal) of the entire dataset with all characters included, (*ii*) removal of a single character and determining whether signal (tRASA) increased or decreased, (*iii*) replacing the removed character and repeating step *ii* with a previously unremoved character, and (*iv*) after steps *ii* and *iii* have been performed for each character in the dataset, excluding those sites whose removal results in increased signal for the remainder of the dataset.

PAUP*4.0b3 (49) was used for tree estimation. All sequences were aligned manually and any parts of the six- and nine-gene matrices with >50% missing data, ambiguous alignments, intron coconversion sites [*coxI* only (25)], and mt RNA edit sites (50) were removed before analyses. Inclusion of ambiguous alignment sites did not fundamentally alter the conclusions drawn in this study. The aligned sequences are available from C.W.D. upon request. Parsimony analyses were performed with two weighting schemes including unweighted (UW) and transversion (T) parsimony. Neighbor-joining (NJ) analyses were conducted using the minimum evolution optimality criterion, assuming different models of nucleotide substitution, including *p* distance (P), Hasegawa–Kishino–Yano (HKY), general-time-reversible (GTR), and GTR + gamma (Γ) = 0.5. Maximum-likelihood (ML) analyses were performed assuming two models of nucleotide substitution, Felsenstein (F81) and HKY. All parameters were simultaneously estimated by using ML. All heuristic searches were performed with 10 random addition sequences and tree bisection reconnection (TBR) swapping. Parsimony and NJ bootstrap analyses were conducted with 500 resampled datasets, whereas ML analyses used only 100.

Results

Single copies of *coxI* and *atpA* were obtained from all species except *Amborella trichopoda*, which had two copies of *atpA*. Fig. 1*A Middle* shows a 9-bp window of sequencing results for total PCR product of *atpA* from *Amborella*. Within this 9-bp region, six sites showed underlying polymorphism suggesting that there is more than one copy of *atpA* in *Amborella*. Cloning experiments recovered one copy that was highly divergent from all others, even at highly conserved sites (Fig. 1*A Top*), whereas a second copy was similar to all other taxa sampled (Fig. 1*A Bottom*). Overall, the two *Amborella atpA* sequences differ at 57 sites. The sequence of the diverged paralog was identical to that reported earlier (12), and as previously noted (12), analyses of this *atpA* sequence suggested *Amborella* was not even a basal angiosperm. In fact, our analyses of 156 *atpA* sequences from a broad

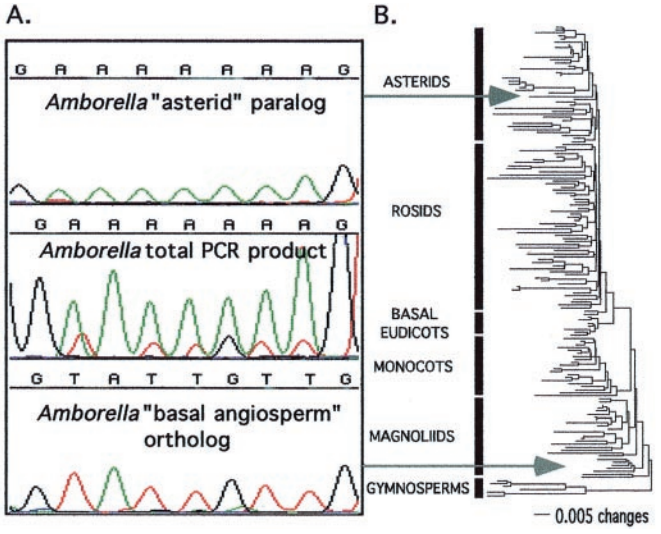

Fig. 1. (*A*) Four-color traces for base pairs 113–121 of two clones and total PCR product for *Amborella atpA*. *Top* shows sequence obtained from 18 clones of the total PCR product shown in *Middle*. *Bottom* shows sequence from one clone obtained from the total PCR product. (*B*) NJ analyses clearly revealed that the sequence in *Bottom* of *A* is orthologous and groups with the basal-most angiosperms, whereas the sequence in *Top* of *A* was likely a paralog of all other basal angiosperm *atpA* sequences, grouping here with the asterid lineage.

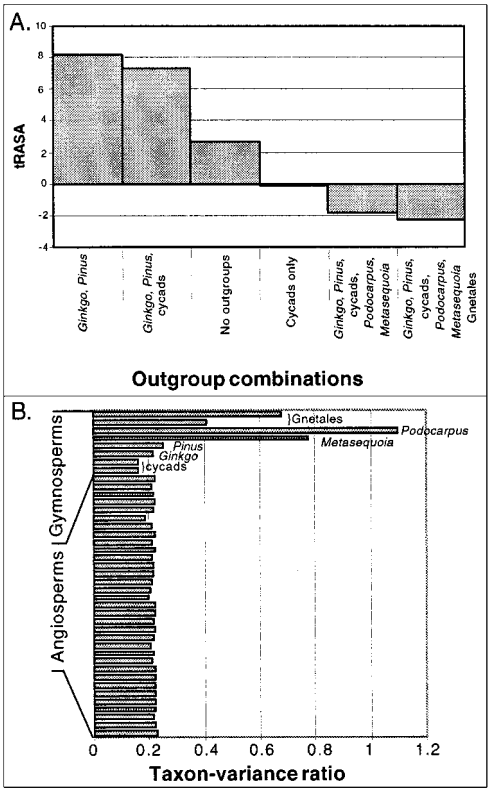

Fig. 2. (*A*) Optimal outgroup analyses (43) using the six-gene dataset indicate that *Ginkgo* and *Pinus* are the best species to use to identify the root node of angiosperm phylogeny. (*B*) Taxon-variance ratio plot showing that Gnetales (*Gnetum* and *Welwitschia*), *Podocarpus*, and *Metasequoia* are outliers relative to all other sampled taxa. Taxon-variance outliers have been shown to lead to inconsistent tree estimation when included in phylogenetic analyses (44).

sampling of angiosperms and gymnosperms placed the *Amborella* paralog within the lineage of higher asterids, whereas the second, putatively orthologous, copy of *atpA* was placed among basal angiosperms as predicted (Fig. 1*B*). For all subsequent analyses, the putative *Amborella atpA* ortholog was used. Importantly, the topology estimated for 156 *atpA* sequences matches estimates from other datasets for the same taxa, suggesting an absence of widespread paralogy (Fig. 1*B*).

Before tree estimation, the six-gene dataset was analyzed with RASA. Optimal outgroup analysis results (Fig. 2*A*) showed that when *Gingko* and *Pinus* were specified as outgroups, signal was highest (tRASA = 8.22), suggesting they are the best taxa with which to infer the root node of tree estimates. All other gymnosperms were removed from subsequent analyses because they were predicted by RASA to reduce the chance of correctly rooting phylogenetic estimates (43). Importantly, signal was higher when *Ginkgo* and *Pinus* were used as outgroups than with no outgroup specified (tRASA = 2.67), indicating that these taxa should not compromise placement of the root. Not only were several gymnosperms poor outgroup choices (gnetophytes, *Podocarpus*, and *Metasequoia*), they were extreme taxon-variance outliers (Fig. 2*B*). This feature suggests their inclusion could have caused parsimony inconsistent estimation conditions (44). None of the in-group species were predicted to be long-branch taxa (Fig. 2*B*). After suboptimal outgroups were removed, RASA 2.4 was used to perform noise reduction of the six-gene combined and individual genomic compartment matrices. The total numbers of noisy characters removed from each dataset may be found at http://depcla4.bio.psu.edu/basals. Final phylogenetic analyses included 35 taxa that, on the basis of RASA data-exploration analyses, were not predicted to compromise phylogenetic estimation.

Combined analyses of the six-gene dataset using NJ (both before and after noise reduction) resulted in branching relationships (Fig. 3) that were largely in agreement with previous studies (11–14), although one striking conflict involving *Amborella* and Nymphaeales will be discussed more below. All methods of analysis estimated major flowering plant lineages, including monocots, Magnoliales, Laurales, Piperales, Eudicots, Winterales, Illiciales, and Nymphaeales, with moderate to high bootstrap support. Also supported was the monophyly of Magnoliales + Laurales, Winterales + Piperales, and Laurales + Magnoliales + Winterales + Piperales. The second branch of angiosperm phylogeny, after *Amborella* + Nymphaeales, was a well-supported clade composed of Illiciales, Trimeniaceae, and Austrobaileyaceae that was sister to all other angiosperms. Interrelationships among five lineages, including monocots, Ceratophyllaceae, Chloranthaceae, Eudicots, and Magnoliales + Laurales + Winterales + Piperales were variously resolved but poorly supported in most cases.

Whereas the estimated relationships discussed above were not method dependent, resolution of the first branch of angiosperm phylogeny was sensitive to method of analysis. Analyses of six genes combined (before RASA noise reduction) using NJ (all models), ML (HKY), and T-parsimony estimated *Amborella* + Nymphaeales (root A; Fig. 4*A*) as the basal-most angiosperm lineage, whereas UW and ML (F81) supported *Amborella*-only (root B; Fig. 4*A*) as the basal-most extant angiosperm. Before and after noise reduction, bootstrap support for the two topologies differed depending on method of analysis (Fig. 4*B*). After noise reduction, UW parsimony estimated *Amborella*-only with low bootstrap support, whereas all other methods estimated *Amborella* + Nymphaeales as the first-branching extant angiosperms with moderate to high support (Fig. 4*B*).

Individual genomic datasets were compiled from the six-gene dataset by concatenating individual genes sampled from each compartment (Table 1). Before noise reduction, NJ analyses of

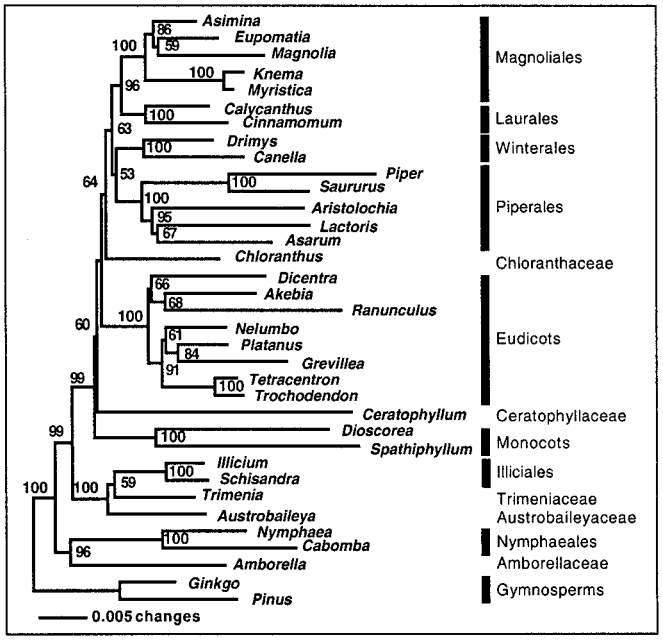

Fig. 3. NJ tree estimated with *p* distances calculated from the noise-reduced dataset of six combined genes. Bootstrap support (BP) is given for nodes >50 and major lineages are labeled. The topology is generally consistent with estimates obtained from raw data and other datasets or methods, except for the first branch of *Amborella* + Nymphaeales (see text).

independent genomes revealed that support for root A (Fig. 4*A*) was greatest with data from the pt and mt, whereas the nuc 18S gene was ambiguous (Fig. 4*C*). UW analyses of the mt genome more strongly supported root A, whereas the pt genome showed marginally stronger support for root B (nuc 18S was ambiguous). After independent genome noise reduction, individual analyses of both organellar genomes revealed stronger support for *Amborella* + Nymphaeales as the first-branching angiosperm lineage when either UW or NJ was used. Analyses of noise-reduced nuc 18S still did not strongly support either root (Fig. 4*C*).

Reanalyses of the 105-taxon dataset (ref. 12; Table 1) showed the same method dependence as found for the six-gene dataset. Analyses including NJ and T parsimony suggested that *Amborella* + Nymphaeales were the basal-most extant angiosperms, whereas UW parsimony strongly suggested *Amborella*-only as reported earlier (12). (It should be noted that taxa missing one or more genes were removed from the NJ analyses to avoid biased distance estimates.)

Analyses of the phytochrome dataset (11) by using NJ suggested that for *PHYA*, *Amborella* is the basal-most extant angiosperm, whereas *Amborella* + Nymphaeales are the first branching lineage for *PHYC* after rooting as previously described (11). In contrast, UW analyses suggested *Amborella*-only was the first-branching angiosperm for *PHYA* and *PHYC* as previously reported (11).

Reanalyses of the 51-taxon dataset (14) revealed the same method dependence as all other analyzed datasets, whereby UW parsimony more strongly supported root B and NJ more strongly supported root A. Although the original published results suggested an *Amborella*-only root with high bootstrap support, it was noted by the authors that an *Amborella* + Nymphaeales root could not be discounted for this dataset (14).

The nine-gene dataset did not include *PHYA* because of the long-branch nature of *Amborella* and the gymnosperm outgroup, *Picea*, as revealed by RASA analyses (see website, http://depcla4.bio.psu.edu/basals). Because these taxa are critical for identifying the angiosperm root and because data exploration suggests that they could mislead phylogenetic estimates, *PHYA* was not considered for this phylogenetic question. Before noise reduction using the optimal outgroup *Pinus*-only, analyses using NJ (all models) and T-parsimony suggested an *Amborella* + Nymphaeales root (Fig. 5 *A* and *B*). UW parsimony and ML analyses (F81 and HKY) resulted in the *Amborella*-only root (Fig. 5*B*). After noise reduction, *all* methods more strongly supported the *Amborella* + Nymphaeales root (Fig. 5*B*). NJ analyses based on mt or nuc data revealed low support for either root, whereas UW more strongly supported *Amborella*-only before noise reduction (Fig. 5*C*). Analyses of pt by using UW or NJ failed to strongly support either root (Fig. 5*C*). After noise reduction of individual genomic compartments, UW estimates of mt and nuc sequences more strongly supported root A, but pt analyses failed to support either root (monocots were the first-branching lineage). All genomes more strongly supported *Amborella* + Nymphaeales when NJ was used on noise-reduced data (Fig. 5*C*).

Discussion

Phylogenetic analyses presented in this paper, based on six and nine genes, strongly suggest that the first-branching extant

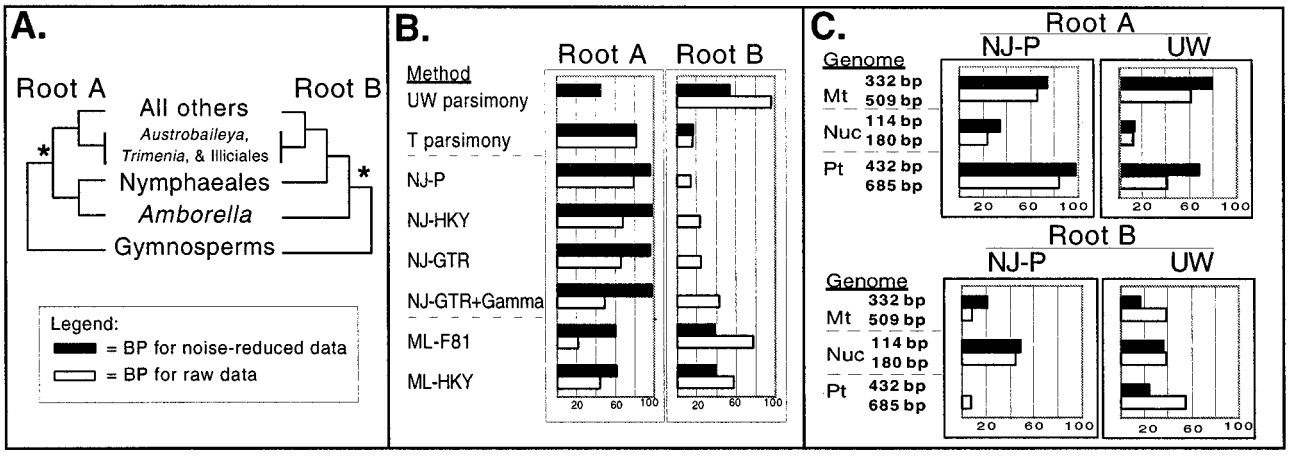

Fig. 4. (A) Alternative roots (marked by *) for basal angiosperm phylogeny based on six-gene analyses. (*B*) Comparisons of BP for estimates of roots A and B in *A,* using multiple methods of analysis from the six-gene combined dataset before and after noise reduction. (*C*) Comparison of BP for both NJ-P and UW parsimony estimates of roots A and B by using individual genomic compartments from the six-gene dataset before and after noise reduction (number of shared variable sites is given in bp).

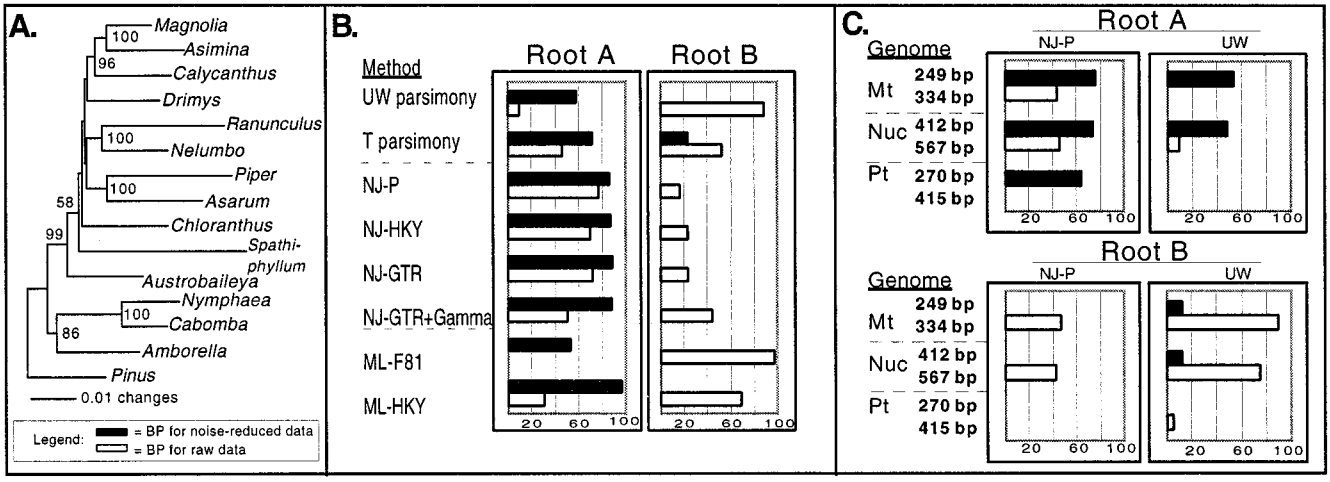

Fig. 5. (*A*) NJ-P tree estimated by using the nine-gene combined noise-reduced dataset. (*B*) Comparison of method-dependent bootstrap support (BP) for the two alternative roots of basal angiosperm phylogeny (see Fig. 4*A*) the nine-gene combined dataset before and after noise reduction was used. (*C*) Comparison of bootstrap support for the two roots when independent genomic compartments before and after noise reduction were used.

angiosperm lineage is composed of *Amborella* + Nymphaeales (Figs. 3 and 5*A*). Confidence in this result is high because (*i*) congruent estimates were obtained by most methods of analysis from combined and individual mt, pt, and nuc genome datasets particularly after noise reduction (only 18S by itself was ambiguous), and (*ii*) high BPs were obtained by assuming various models from both independent and combined analyses. Furthermore, the Kishino–Hasegawa test (51) revealed that root A was a significantly better fit to the data ($P < 0.05$) than root B for the nine-gene dataset (six-gene dataset was not significant at $P < 0.05$). Previously, confidence was assumed for the first branching position of *Amborella*-only because of the high bootstrap values obtained and the fact that multiple studies came to the same result (11–14). These individual studies lacked congruence as a measure of reliability because the analyses only used UW parsimony (except ref. 14, which also used one ML analysis), and when independent genomes were considered, incongruence was noted (12, 14). The major difference between the two topologies is placement of the root, one node apart, because analyses excluding all gymnosperms find an essentially identical topology with all methods and genomes (results not shown). The results obtained previously were shown to be method dependent also; thus, the major conclusions obtained from our study are not due to reduced taxon or additional gene sampling. There are evolutionary implications for accepting either phylogenetic estimate (discussed below); therefore, it is of importance to consider the biological and methodological factors that affected the topological estimates obtained.

Paralogy, Outgroups, and Long-Branch Taxa. The existence of multiple loci for *atpA* in *Amborella* is surprising, although it has been reported in *Oenothera* (52). The inclusion of the highly diverged mt *atpA* paralog (12) significantly affected inference of phylogeny using *atpA* or mt sequences only and compromised congruence among genomic estimates (results not shown). Accordingly, we included only the *Amborella* ortholog in the final analyses.

As shown above, not only were some gymnosperm outgroups suboptimal choices for identifying the root of angiosperm phylogeny (Fig. 2*A*), but taxon-variance plots indicated that their inclusion could cause UW parsimony inconsistent conditions (Fig. 2*B*). As evidence of this, use of long-branch outgroups for the pt-only dataset resulted in a well-supported root node with *Ceratophyllum* estimated as the first-branching angiosperm (results not shown), a result incongruent with most recent molecular phylogenies (refs. 12–14 and Fig. 3). However, it should be noted that analyses of total combined data by using suboptimal outgroups did not fundamentally alter the conclusions of this study.

Furthermore, the RASA-based identification of *Amborella PHYA* and *Picea* phytochrome as "long branches" allowed us to avoid obtaining potentially spurious tree estimates when analyzing this gene. Indeed, since 1978 it has been recognized that some sequences or taxa can cause incorrect phylogenetic estimation (33). Now that problematic taxa can be identified by using objective criteria (44), they can be removed (or examined) if their inclusion is predicted to mislead phylogenetic estimates.

Methodological and Genomic Congruence. In our analyses, RASA-based data exploration tended to reduce the apparent conflict in the six- and nine-gene datasets. Removal of phylogenetic noise led to a dramatic increase of congruence among genomes and methods. Because all datasets are likely to have some form of phylogenetic noise, suboptimal outgroups, and long-branch taxa, methods such as RASA are becoming critical parts of many rigorous analyses (7, 45–48).

Congruence (19, 20) can indicate reliability among phylogenetic estimates when evolutionary history is not known. However, congruence among standard phylogenetic methods may be unattainable for the subsets of basal angiosperms and gymnosperm outgroups analyzed in this paper. As shown in Fig. 6, UW shows decreasing support for root A with increasing character sampling of raw data (increasing support for root B). On the other hand, NJ analyses or UW with noise-reduced data show increasing support for root A with increased character sampling. The trend implied by Fig. 6 suggests that if 10 times as much data (of the same quality) were available for this group of taxa, UW would unequivocally support root B (BP = 100), whereas NJ and UW with noise reduction would unambiguously support root A (BP = 82–100). This result implies that the bootstrap alone cannot be used to indicate reliability (41); instead, congruent genomic estimates that are well supported may be a better indicator of reliable results.

Implications. Does it matter that the first-branching angiosperm lineage may well be composed of *Amborella* + Nymphaeales rather than *Amborella* only? There are implications to accepting any phylogenetic estimate, particularly when the branching order affects inferences and generalizations about character evolution.

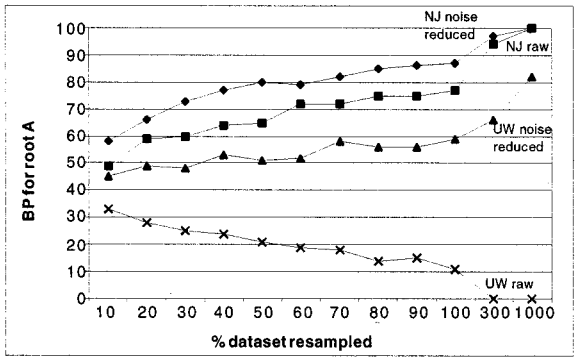

Fig. 6. Bootstrap power curve showing BP support for root A (Fig. 4*A*) when NJ and UW are used on both the raw and noise-reduced nine-gene dataset. Various proportions of the nine-gene dataset were resampled (10–1000% of the original data) to investigate the effect of numbers of characters on BP for root A (Fig. 4*A*). Note that if 10 times (1000%) the amount of data (of the same quality) were sampled, NJ and UW (using noise-reduced data) would unambiguously support root A, whereas UW (using raw data) would provide no support for this root.

For instance, if *Amborella* is assumed to be the first-branching angiosperm, parsimony reconstructions of plant sexuality (based on extant taxa) suggest that the ancestor of angiosperms had unisexual flowers; however, if *Amborella* + Nymphaeales is assumed to be the first-branching angiosperm lineage, then the ancestor of angiosperms may have had either unisexual or bisexual flowers. Furthermore, the ancestor of angiosperms is unequivocally vesselless in a topology that assumes *Amborella*-only, but if *Amborella* + Nymphaeales is the basal-most lineage, then either the ancestor of all angiosperms had vessels and *Amborella* lost them, or vessels have evolved on more than one occasion in angiosperm history. The latter hypothesis supports detailed anatomical observations of conducting elements in the Nymphaeales that show an incipient stage of vessel evolution (53, 54). Other inferences about the most recent common ancestor of angiosperms, such as the woody growth habit, and origin in tropical Gondwanaland, do not differ when assuming either of the two topologies.

It appears that most problems known to decrease the accuracy of phylogenetic reconstruction are present in datasets accumulated for basal angiosperms. Because of the complicated nature of this tree estimation problem, only congruence among genomes and methods will result in high confidence in phylogenetic estimates. Although more data of all kinds will likely be collected to study basal angiosperm relationships (55), it is clear that detailed analyses should be a critical component of future studies.

We thank Ned Young for several *cox1* sequences and Lena Landherr for assistance; Susanne Renner, the Soltis laboratory, the University of California Santa Cruz, the Missouri Botanical Garden, the University of California Berkeley Botanical Garden, and Joe Armstrong for DNA or plant material; Jeffrey Palmer, Sarah Mathews, Y.-L. Qiu, Pamela and Douglas Soltis, and Mark Chase for making their published datasets available; Blair Hedges, Michael Donoghue, Sarah Mathews, and Marcel van Tuinen for helpful suggestions; and Penn State University and the National Science Foundation for financial support.

1. Darwin, C. (1903) in *More Letters of Charles Darwin: A Record of His Work in a Series of Hitherto Unpublished Letters*, eds. Darwin, F. & Seward, A. C. (John Murray, London), Vol. 2.
2. Crane, P. R., Fris, E. M. & Pedersen, K. R. (1995) *Nature (London)* **374,** 27–33.
3. Doyle, J. A. (1998) *Mol. Phylogenet. Evol.* **9,** 448–462.
4. Carmichael, J. S. & Friedman, W. E. (1996) *Am. J. Bot.* **83,** 767–780.
5. Winter, K.-U., Becker, A., Munster, T., Kim, J. T., Saedler, H. & Theissen, G. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 7342–7347.
6. Hansen, A., Hansmann, S., Samigullin, T., Antonov, A. & Martin, W. (1999) *Mol. Biol. Evol.* **16,** 1006–1009.
7. Bowe, L. M., Coat, G. & dePamphilis, C. W. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4092–4097.
8. Chaw, S.-M., Parkinson, C. L., Cheng, Y., Vincent, T. M. & Palmer, J. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4086–4091.
9. Hughes, N. F. (1994) *The Enigma of Angiosperm Origins* (Cambridge Univ. Press, Cambridge, U.K.).
10. Crepet, W. L. (1998) *Science* **282,** 1653–1654.
11. Mathews, S. & Donoghue, M. J. (1999) *Science* **286,** 947–950.
12. Qiu, Y.-L., Lee, J., Bernasconi-Quadroni, F., Soltis, D. E., Soltis, P. S., Zanis, M., Zimmer, E. A., Chen, Z., Savolainen, V. & Chase, M. W. (1999) *Nature (London)* **402,** 404–407.
13. Soltis, P. S., Soltis, D. E. & Chase, M. W. (1999) *Nature (London)* **402,** 402–404.
14. Parkinson, C. L., Adams, K. L. & Palmer, J. D. (1999) *Curr. Biol.* **9,** 1485–1488.
15. Wolfe, K. H., Li, W.-H. & Sharp, P. M. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 9054–9058.
16. Palmer, J. D. (1990) *Trends Genet.* **6,** 115–120.
17. Morgensen, H. L. (1996) *Am. J. Bot.* **83,** 383–404.
18. Gray, M. W., Burger, G. & Lang, F. B. (1999) *Science* **283,** 1476–1481.
19. Miyamoto, M. M. & Cracraft, J. (1991) in *Recent Advances in Phylogenetic Analysis of DNA Sequences*, eds. Miyamoto, M. M. & Cracraft, J. L. (Oxford Univ. Press, Oxford, U.K.), pp. 3–17.
20. Miyamoto, M. M. & Fitch, W. M. (1995) *Syst. Biol.* **44,** 64–76.
21. Pamilo, P. & Nei, M. (1988) *Mol. Biol. Evol.* **5,** 568–583.
22. Syvanen, M. (1994) *Annu. Rev. Genet.* **28,** 237–261.
23. Chase, M. W. & Albert, V. A. (1998) in *Molecular Systematics of Plants II: DNA Sequencing*, eds. Soltis, D. E., Soltis, P. S. & Doyle, J. J. (Kluwer, Boston), pp. 488–507.
24. Nandi, O. I., Chase, M. W. & Endress, P. K. (1998) *Ann. Missouri Bot. Garden* **85,** 137–212.
25. Cho, Y., Qiu, Y.-L., Kuhlman, P. & Palmer, J. D. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14244–14249.
26. Sanderson, M. J. & Doyle, J. J. (1992) *Syst. Biol.* **41,** 4–17.
27. Kim, J. (1993) *Syst. Biol.* **42,** 331–340.
28. Saitou, N. & Imanishi, M. (1989) *Mol. Biol. Evol.* **6,** 514–525.
29. Huelsenbeck, J. P. (1995) *Syst. Biol.* **44,** 17–48.
30. Hillis, D. M., Bull, J. J., White, M. E., Badgett, M. R. & Molineux, I. J. (1992) *Science* **255,** 589–592.
31. Tateno, Y., Takezaki, N. & Nei, M. (1994) *Mol. Biol. Evol.* **11,** 261–277.
32. Nei, M., Kumar, S. & Takahashi, K. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 12390–12397.
33. Felsenstein, J. (1978) *Syst. Zool.* **27,** 401–410.
34. Takezaki, N. & Nei, M. (1994) *J. Mol. Evol.* **39,** 210–218.
35. Kim, J. (1996) *Syst. Biol.* **45,** 363–374.
36. Zharkikh, A. & Li, W.-H. (1993) *Syst. Biol.* **42,** 113–125.
37. Kuhner, M. K. & Felsenstein, J. (1994) *Mol. Biol. Evol.* **11,** 459–468.
38. Yang, Z. (1996) *J. Mol. Evol.* **42,** 294–307.
39. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
40. Efron, B., Halloran, E. & Holmes, S. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 7085–7090.
41. Nei, M. (1996) *Annu. Rev. Genet.* **30,** 371–403.
42. Lyons-Weiler, J., Hoelzer, G. A. & Tausch, R. J. (1996) *Mol. Biol. Evol.* **13,** 749–757.
43. Lyons-Weiler, J., Hoelzer, G. A. & Tausch, R. J. (1998) *Biol. J. Linn. Soc.* **64,** 493–511.
44. Lyons-Weiler, J. & Hoelzer, G. A. (1997) *Mol. Phylogenet. Evol.* **8,** 375–384.
45. Culligan, K. M., Meyer-Gauen, G., Lyons-Weiler, J. & Hays, J. B. (2000) *Nucleic Acids Res.* **28,** 463–471.
46. van Tuinen, M., Sibley, C. G. & Hedges, S. B. (2000) *Mol. Biol. Evol.* **17,** 451–457.
47. Wolf, P. G. (1997) *Am. J. Bot.* **84,** 1429–1440.
48. Teeling, E. C., Scally, M., Kao, D. J., Romagnoli, M. L., Springer, M. S. & Stanhope, M. J. (2000) *Nature (London)* **403,** 188–189.
49. Swofford, D. L. (1998) PAUP*, *Phylogenetic Analysis Using Parsimony (and Other Methods)* (Sinauer, Sunderland, MA).
50. Bowe, L. M. & dePamphilis, C. W. (1996) *Mol. Biol. Evol.* **13,** 1159–1166.
51. Kishino, H. & Hasegawa, M. (1989) *J. Mol. Evol.* **29,** 170–179.
52. Schuster, W. & Brennicke, A. (1986) *Mol. Gen. Genet.* **204,** 29–35.
53. Carlquist, S. (1996) in *Flowering Plant Origin, Evolution and Phylogeny*, eds. Taylor, D. W. & Hickey, L. J., (Chapman & Hall, New York), pp. 68–90.
54. Schneider, E. L. & Carlquist, S. (1996) *Am. J. Bot.* **83,** 1236–1240.
55. Graham, S. W. & Olmstead, R. G., *Am. J. Bot.,* in press.