Syst. Biol. 62(1):50–61, 2013

DOI:10.1093/sysbio/sys070
Advance Access publication on July 31, 2012

The Evolutionary Root of Flowering Plants

VADIM V. GOREMYKIN[1]*, SVETLANA V. NIKIFOROVA[1], PATRICK J. BIGGS[2], BOJIAN ZHONG[3], PETER DELANGE[4], WILLIAM MARTIN[5], STEFAN WOETZEL[6], ROBIN A. ATHERTON[3], PATRICIA A. MCLENACHAN[3], AND PETER J. LOCKHART[3,7]

[1]IASMA Research Center, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy; [2]Institute of Veterinary, Animal and Biomedical Sciences, Massey University, Palmerston North, New Zealand; [3]Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand; [4]Department of Conservation, Auckland Conservancy, New Zealand; [5]Institut für Botanik III Heinrich-Heine-Universität, Germany; [6]Max Planck Institute for Plant Breeding Research, Cologne, Germany; and [7]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand
**Correspondence to be sent to: IASMA Research Center, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy;*
E-mail: Vadim.Goremykin@iasma.it.

Received 14 December 2011; reviews returned 21 February 2012; accepted 26 July 2012
Associate Editor: Lars Jermiin

Abstract.—Correct rooting of the angiosperm radiation is both challenging and necessary for understanding the origins and evolution of physiological and phenotypic traits in flowering plants. The problem is known to be difficult due to the large genetic distance separating flowering plants from other seed plants and the sparse taxon sampling among basal angiosperms. Here, we provide further evidence for concern over substitution model misspecification in analyses of chloroplast DNA sequences. We show that support for *Amborella* as the sole representative of the most basal angiosperm lineage is founded on sequence site patterns poorly described by time-reversible substitution models. Improving the fit between sequence data and substitution model identifies *Trithuria*, Nymphaeaceae, and *Amborella* as surviving relatives of the most basal lineage of flowering plants. This finding indicates that aquatic and herbaceous species dominate the earliest extant lineage of flowering plants. [*Trithuria inconspicua*; chloroplast genome; angiosperm origins; heterotachy; base compositional heterogeneity; data model fit.]

Although there is increasing consensus about many relationships among major lineages of flowering plants (Soltis et al. 2011) and convergence toward more similar dates for the origin of angiosperms (Jiao et al. 2011; Sun et al. 2011), determining the root of the angiosperm phylogeny has been more problematic. This difficulty is not unique to the study of angiosperms; reconstructing basal relationships in species radiations is known to be hard (Shavit et al. 2007; Graham and Iles 2009). Not only can the shape of the true underlying phylogeny make it difficult to accurately reconstruct gene trees (Whitfield and Lockhart 2007), even correct gene trees can be incongruent with the underlying species phylogeny (Degnan and Rosenberg 2009).

In phylogenetic studies of chloroplast DNA (cpDNA), nuclear DNA (nuDNA), and mitochondrial DNA (mtDNA), *Amborella* has often been recovered as the sole survivor of the first lineage to diverge from that leading to all the other extant flowering plants (Mathews and Donoghue 1999; Qiu et al. 1999; Soltis and Soltis 2004; Stefanovié et al. 2004; Leebens-Mack et al. 2005; Jansen et al. 2007; Saarela et al. 2007; Graham and Iles 2009; Soltis et al. 2011). However, a closer relationship between *Amborella* and aquatic angiosperm species has been reported in analyses of mitochondrial and nuclear DNA (Qiu et al. 2010; Jiao et al. 2011; Soltis et al. 2011;) as well as in model-based analyses of chloroplast genes that typically exclude or reduce the impact of third codon positions (Barkman et al. 2000; Wu et al. 2007). Opinion has been divided over how to treat third codon positions in cpDNA. Although inclusion of these sites might improve phylogenetic resolution between some taxa (Zanis et al. 2002; Stefanovié et al. 2004; Leebens-Mack et al. 2005), they also exhibit evidence of a decayed historical signal (due to multiple substitutions at the same site) between some taxa (Goremykin et al. 2003; Chaw et al. 2004). Analyses of short independent nuclear markers have not provided improved phylogenetic resolution, suggesting instead alternative relationships among basal angiosperms (e.g., Mathews and Donoghue 1999; Jiao et al. 2011; Soltis et al. 2011). This finding is perhaps not unexpected given the short internal branches typically reconstructed for angiosperm phylogenies (e.g., see Martin et al. 2005).

We have previously suggested that a poor fit between commonly used phylogenetic models and sequence data contributes to uncertainty concerning relationships among early diverging lineages of flowering plants (Lockhart and Penny 2005; Martin et al. 2005). Here, we provide further evidence for this hypothesis in a study of the substitution properties of concatenated chloroplast genome sequences, and in particular of sites in the alignment that are most varied. These sites, often called "fast sites," show the greatest character state variation as well as evidence of multiple substitutions. Numerous methods for sorting, identifying, and removing fast sites have been suggested, and the impact of removal of the fastest evolving sites on phylogenetic reconstruction is well known (e.g., Brinkmann and Philippe 1999; Hirt et al. 1999; Lopez et al. 1999; Ruiz-Trillo et al. 1999; Hansmann and Martin 2000; Burleigh and Mathews 2004; Pisani 2004). Less well appreciated is the observation that the sorting of sites based on character state variation or compatibility criteria allows the properties of sites that impact on tree building to be more easily studied (Sperling et al. 2009). We have examined the compositional heterogeneity of fast sites and the fit of concatenated chloroplast sequences to the GTR+I+Γ substitution model commonly used in angiosperm phylogeny studies. We address the problem of identifying which of the fast sites to exclude from the phylogenetic data by applying the GNB criterion (named after the inventors: Goremykin et al. 2010) to the concatenated alignment after the sites in this alignment had been reordered according to their observed variability (OV; see "Materials and Methods"). This criterion has been suggested as suitable for identifying sites most affected by multiple substitutions in a multiple sequence alignment. Here, we examine the properties of the fast sites identified under the GNB criterion and the contribution of these sites to topological distortion in phylogenetic trees reconstructed for angiosperm and conifer sequences. To obtain optimal phylogenetic estimates, we employed the CAT+covarion model, which was consistently identified in our cross-validation analyses as the best-fitting model to our original data and to data partitions generated in the "noise reduction" protocol of Goremykin et al. (2010). This substitution model better accommodates a restricted substitution profile across sites and describes spatial heterogeneity of substitutions in terms of simple covarion models (Ane et al. 2005).

To improve taxon sampling at the base of the angiosperm radiation, we also sequenced the chloroplast genome of *Trithuria inconspicua*, a species from a genus of minute aquatic herbs, which recently has been found to be closely related to Nymphaeaceae (Saarela et al. 2007). Our findings highlight the importance of the fit between model and data when evaluating relationships among basal angiosperms.

Materials and Methods

Sequencing of the Chloroplast Genome of *Trithuria inconspicua*

Trithuria inconspicua was collected from the Kai Iwi Lakes (Lakes Waikare and Taharoa), Northland, North Island, New Zealand, and sent by courier to Massey University, Palmerston North. Voucher specimens have been deposited at the Auckland War Memorial Museum Herbarium AK (see AK 308938, AK 320388). Enriched cpDNA was sequenced on an Illumina GAII platform as described in Atherton et al. (2010). Contigs were assembled using Velvet version 0.7.60 (Zerbino and Birney 2008) and odd kmer values ranging from 25 to 61. Because the copy number of cpDNA was higher than that for the nuDNA (though not a higher absolute amount), coverage cutoffs of 10, 20, 40, and 80 were applied during the assembly of contigs. Staden 2.0.0b7 (http://staden.sourceforge.net/) was used to join the contigs generated by Velvet. Nine gaps remained after the assembly; 8 gaps were closed by designing primers to flanking regions and sequencing the missing parts using standard ABI3730 sequencing protocols (Massey Genome Service http://genome.massey.ac.nz/).

Taxon Selection and Multiple Sequence Alignment

Protein-coding sequences of 61 genes common to 31 chloroplast genomes from angiosperms and gymnosperms were downloaded from GenBank. NAD dehydrogenase genes were not included in analyses as these are absent from the cpDNAs of gnetophytes and conifers (Wakasugi et al. 1994; Braukmann et al. 2009). In our taxon sampling, we included representatives of all available basal angiosperm lineages but not all crown group angiosperm species for which chloroplast genomes have been determined. This taxon selection retained species most important for inferring relationships among basal angiosperms and reduced computation time for model-fitting and tree-building analyses on a 16-core Linux server. Eudicots were represented by 6 basal species. We excluded grasses, known to be subtended by a very long branch in previous analyses (Goremykin et al. 2005), keeping all other monocots.

As concern over alignment procedures remains an important practical consideration for phylogenomic analyses (Philippe et al. 2011), multiple sequence alignments were generated using 2 alignment protocols in the present study. The first protocol, used as a basis for figures shown in this article, uses the same principles described in Goremykin et al. (2004). This alignment protocol provides a rapid and reliable method of aligning similar gene sequences and for producing data sets comprising first and second codon positions and all 3 codon positions. With this approach, gene sequences were sorted into 61 Fasta files, each containing orthologs. For each file, first and second codon positions were aligned using the program MUSCLE (Edgar 2004). Alignments for sequences that included all 3 codon positions were also generated by the same script. The resulting 122 alignment files were each manually edited, such that regions of low similarity between the ingroup and outgroup sequences were discarded. Individual gene alignment files were concatenated using Geneious v5.5.4. (Drummond et al. 2010) to produce: (i) a gapped alignment of 40 553 positions in length, provided as a supplementary material (Supplementary File S1) and (ii) an alignment of first and the second codon positions 25 246 positions in length (Supplementary File S2). An OV sorted (see below) version of the 40 553 pos. long alignment has been provided as a Supplementary File S3.

A MUSCLE alignment of translated nucleotide sequences from 56 individual Fasta files was also generated and used to confirm results of phylogenetic analyses obtained using the first alignment protocol. This second alignment approach used the same principles as previously implemented for obtaining conservative alignments between anciently diverged sequences (Lockhart et al. 1996). With this method, we imported each Fasta file into MEGA 5.0 (Tamura et al. 2011), translated the sequences, and then aligned them with MUSCLE (default options). We concatenated these aligned files using Geneious v5.5.4. (Drummond et al. 2010) and then imported the concatenated file into Se-Al.

v2.0a11. (Rambaut 2002). Site patterns adjacent to indels were then removed if they did not contain amino acids with similar physical/chemical properties as specified in Se-Al. Finally, the columns with gaps were removed and the sequences back-translated. This alignment protocol produced a much shorter concatenated alignment than did the first method (31 674 ungapped positions). This alignment has been provided as a supplementary material (Supplementary File S4).

OV Sorting and "Noise Reduction"

Site patterns in our concatenated alignments were reordered according to their OV scores and data partitions identified for tree building using the GNB criterion (Goremykin et al. 2010). Previously, this approach was found effective in the recovery of benchmark clades of mammalian phylogeny, and more effective than other methods in identifying fast-evolving sites that cause long branch attraction (LBA) artifacts (Goremykin et al. 2010).

OV sorting involves calculating a sum-of-pairs mismatch score for each site in the full alignment (including positions with gaps) and then ordering the sites according to the OV scores (Goremykin et al. 2010). This produces an alignment with the most conserved (least varied) site patterns at one end, and the least conserved (most varied) positions at the other end. We refer to this alignment as the OV alignment. The OV alignment was generated using the script Sorter.pl. This script also splits the OV alignment into several bipartitions of sites. Each bipartition contains an "A" partition, which includes site patterns from the conserved end of the alignment, and a "B" partition, which includes site patterns from the least conserved end of the OV alignment. In the present study, the bipartition of sites into partitions A and B occurred at position $i \times 250$ (where $i = 1, 2, 3, \ldots$) upstream from the most varied end of the OV alignment. The incremental increase in interval length of 250 sites for the B partition is an arbitrary size previously found suitable for monitoring change in the properties of the ordered sites at the most varied end of the OV alignment. Once the bipartitions are formed, the script Sorter.pl calls ModelTest (Posada and Crandall 1998) to identify an optimal time-reversible substitution model for each of the A and B partitions using a 2-step procedure (for further details, see Goremykin et al. 2010). The script then calls PAUP* (Swofford 2002; Unix v. 4.0b10) to calculate a matrix of maximum likelihood (ML) distances for the A and B partitions. A matrix of p-distances (number of sites with observed differences/total number of sites) is also calculated for each B partition.

Sorter.pl also calculates the average of the ML-distances minus the average of the p-distances and reports this mean deviation of the ML- and p-distances for the B partitions, and Pearson correlation coefficient values (r) between these estimates (Goremykin et al. 2010). Dissimilarity between relative ranking of ML- and p-distances calculated from the B partitions occurs if distance estimates are not similar between taxa. Stochastic error associated with the short sequence length of the initial B partitions will cause such dissimilarity, as will substitution model violations and saturation with multiple substitutions. By monitoring the r values as the length of the B partition is increased, it is possible to identify a point of transition with respect to the similarity of the distances compared. As the relative ranking of absolute distance values within 2 groups of distance estimates (p- and ML-distances) becomes similar, there is a dramatic rise in the value of r.

In addition to comparing the ML- and p-distances for B partitions, the script Sorter.pl also compares optimal ML-distances for the A and B partitions. Deviation is again measured in terms of r. As with the ML- and p-distance comparison for the B partition, a dramatic rise in r occurs when the distances become proportional, and their ranking becomes similar. The comparison identifies the relative length of the A and B partitions, at which point the evolutionary properties of the B partition become similar to those of the A partition.

Goremykin et al. (2010) have suggested that the site stripping process should cease when there is a dramatic increase in the value of r in both correlation analyses. At this point, positions added from the conserved A partition to the variable B partition clearly start to mask the nonphylogenetic signal associated with the most varied positions in the B partitions. Here, we also report that the topological distortion induced by the presence of B partition sites is also greatly reduced at this point. Model misspecification contributed by compositional heterogeneity, as we also show, still persists beyond this point. However, this has little impact on the relative ranking of distances in B partitions. Thus, further character removal is not justified on the basis of the GNB criterion.

As demonstrated in Zhong et al. (2011), the GNB criterion also identifies and provides a basis for removing sites from a concatenated alignment that have a poor fit to phylogenetic model assumptions. Although this criterion does not remove all model-violating sites from the data, it has been shown to remove sites that significantly impact on phylogenetic estimates, and thus sites that have significant effect in misleading tree building. In particular, it appears very useful for reducing LBA artifacts in phylogenetic reconstruction. This was demonstrated in reanalysis of mitochondrial DNA sequences, which previously and consistently had yielded a rodent polyphyly artifact (Goremykin et al. 2010) and also in recent analyses of chloroplast sequences from Gnetales and other seed plants (Zhong et al. 2011).

To study the relationship between changes in r and branch length support in reconstructed trees, splits can be calculated for individual A and B partitions. We calculated NeighborNet (NNET: Bryant and Moulton 2004) splits from the optimal ML-distances obtained for each B partition generated during the noise reduction protocol. These were calculated using SplitsTree 4.0 (Huson and Bryant 2006). Of particular interest are the splits that separate outgroup and ingroup taxa as these are relevant for the question of rooting the angiosperm radiation. In the present study, we plotted the relative size of the splits separating: (i) angiosperms from gymnosperms and (ii) Gnetales from other species. Such a "heterotachy plot," as it was referred to in Zhong et al. (2011), allows visualization of the relationship between B-partition distances and any topological distortion (Lockhart et al. 1996; Bruno and Halpern 1999) of reconstructed trees due to including the most varied sites of the OV alignment when tree building.

Base Composition Heterogeneity

Base compositional heterogeneity (Lockhart et al. 1992; Jermiin et al. 2004) was examined over the most varied end of the OV alignment. To investigate this, intervals of sites with the same length (360 jacknife resampled ungapped positions; 3 replicates for each interval) were sampled from nonoverlapping locations at the most varied end of the OV alignment (between 0 and 500 sites, 500 and 1000 sites, 1000 and 1500 sites, ... , 9500 and 10 000 sites). We examined each of these sets of sites using Bowker's matched-pair symmetry test (Ababneh et al. 2006), as implemented in Seq-Vis (Ho et al. 2006). We used Seqboot from the PHYLIP v3.69 (Felsenstein 2004) package for jackknife resampling of sites (sampling without replacement) and SeqVis v1.5 (Ho et al. 2006) for the symmetry test. The smallest interval from which sites were resampled was the first interval: 0–500 sites (these 500 gapped positions contained 380 ungapped positions).

Goodness of Fit Analyses

We used MISFITS (Nguyen et al. 2011) and Tree-Puzzle-5.2 (Schmidt et al. 2002) to identify those site patterns in the OV alignment whose observed frequencies were unexpected under a GTR+I+Γ substitution model. This model was identified as the best-fitting model to the OV alignment among all models that assumed a single matrix of base frequencies. This was also the case for the increasingly short A partitions according to a double-fitting procedure that employed an Akaike information criterion (AIC) (described in Goremykin et al. 2010). The fit of the GTR+I+Γ model to chloroplast data sets is also of significant interest as this model has been commonly used in phylogenetic analyses of basal angiosperms (e.g., Barkman et al. 2000; Zanis et al. 2002; Stefanović et al. 2004; Leebens-Mack et al. 2005; Saarela et al. 2007; Wu et al. 2007; Graham and Iles 2009; Qiu et al. 2010; Jiao et al. 2011; Soltis et al. 2011). Although there are computational issues with coestimation of the I+Γ parameter values (e.g., see "Discussion" in Yang 2006), this model has been found to have higher reconstruction accuracy than GTR+Γ models in more biologically realistic simulations (Gruenheit et al. 2008). The impact that deletion of sites from the most varied end of the OV alignment had on the fit of this substitution model was also studied at different shortening steps. Log-likelihood scores for the evolutionary model obtained for the increasingly short A partitions were also compared with the log-likelihood scores for equal length partitions that were jackknife resampled from the complete OV alignment. We used Seqboot for jackknife resampling and PhyML 3.0 (Guindon et al. 2010) for calculating log-likelihood scores.

Substitution Model Selection for A Partitions

The optimal substitution model was determined for the A partition data sets using cross-validation as implemented in PhyloBayes 3.2e (Lartillot and Philippe 2004). To determine the length of time needed for convergence of posterior probabilities, we initially ran PhyloBayes on a 16-core Linux server for at least 2 weeks with alignments of the first and second codon positions, and of all 3 codon positions, choosing between 6 substitution models for each input file: The "classical" GTR+Γ, GTR+Γ+covarion, GTR+Γ+covext, GTR+Γ+CAT, GTR+Γ+CAT+covarion, and GTR+Γ+CAT+covext (Lartillot and Philippe 2004). Here, "CAT" refers to the site-heterogeneous mixture model of Lartillot and Philippe (2004), "covarion" to the covarion model of Tuffley and Steel (1998), and "covext" to a variant of the Tuffley and Steel model that allows for variation in rate heterogeneity across sites. We assumed a four-category discrete Γ distribution in modeling rate-heterogeneity across sites. From these initial 12 runs, we determined that 200 cycles were sufficient for convergence on our Linux server. Since cross-validation is multistaged and computationally demanding, we wrote a script Cross.pl, which initiates parallel multiple PhyloBayes and cross-validation runs. This script first invokes PhyloBayes, lets it run for 1000 cycles under the abovementioned models, and builds consensus trees discarding the first 200 cycles as burn-in. Then the script invokes the PhyloBayes program cvrep to randomly sample 10 learning and 10 test data partitions from each alignment, so that each learning data partition has 90% of the input alignment length and each test partition has 10% of the input alignment length. The script then calls PhyloBayes and performs Markov chain Monte Carlo sampling for 200 cycles in parallel for the learning sets created by the PhyloBayes program cvrep. Subsequently, the script initiates the Phylo Bayes program readcv in parallel for all data replicates and computes a cross-validation score (i.e., calculates the likelihood under the test set, averaged over the posterior distribution of the learning set) discarding a burn-in of 50 sampling points and taking every point thereafter. Finally, the script invokes the PhyloBayes program sumcv to compute summary statistics. Using the AIC for the double-fitting procedure, the GTR+I+Γ model was selected as the best-fitting model among those with one matrix of base frequencies for the OV alignment and its next 20 shortened subsets.

Tree Building

Phylogenetic reconstructions were performed using PhyloBayes and the PAUP*-embedded scripts in Sorter.pl (Goremykin et al. 2010). RAxML (Stamatakis et al. 2005) was also used to reanalyze a recently published data set of chloroplast, mitochondrial, and nuclear genes (Soltis et al. 2011).

Availability of scripts and of Trithuria chloroplast genome sequence.—Scripts not already publically available and used in this study have been provided as supplementary material. The sequence for the *Trithuria inconspicua* chloroplast genome determined in this study has been deposited with EMBL (Accession no. HE963749).

Results

Alignments

Two alignments were obtained using different approaches in the present study. Despite differences in their lengths, both methods produced very similar alignments. This can be visualized by comparing split networks that display the NNET split systems (*p*-distances) for each alignment (Supplementary File S5). Similar analytical results were obtained for both alignments. The figures shown in subsequent sections were based on the alignment method of Goremykin et al. (2003).

GNB Analyses

A significant improvement in *r* occurred after 8 steps: 2000 sites (Fig. 1a); that is, once the 2000 most varied sites were included in the B partition, *p*-distances and ML-distances for the B partition had become highly correlated. Similarly, at this shortening step ML-distances for A and B partitions also became highly correlated (Fig 1b), indicating similar evolutionary distances for both partitions, and suggesting a point had been reached at which further removal of sites from the A partition was no longer justified. Most significantly, the distance between outgroup and ingroup taxa reduced dramatically by the eighth sampling step. This was visualized in Figure 2, which shows the relative length of outgroup splits in the NNET split system for the taxon set. The extreme branch length separating the outgroup and ingroup sequences is a property of the 2000 sites at the most varied end of the OV alignment.

Compositional Heterogeneity

It has been previously observed that compositional heterogeneity and the rate of substitutions of sites are tightly correlated (Rodriguez-Ezpeleta et al. 2007). Our analyses provide some support for this observation. Figure 3 indicates that compositional heterogeneity is a feature of the most varied end of the OV alignment. In particular, it indicates the number of pairs failing a matched-pairs test of symmetry at $P < 0.00005$ when these are calculated on identical length partitions (360 sites each) sampled within 500-bp nonoverlapping gaped intervals at the most varied end of the OV alignment. The plot suggests that heterogeneity in composition is most significant over the first 3000–3500 most varied positions of this alignment. This heterogeneity is most significant between angiosperm and outgroup sequences and among outgroups sequences (values for individual pairs not shown). It extends past the stopping point identified by the GNB method. Hence, although compositional heterogeneity is likely to contribute to the extreme branch length difference between ingroup and outgroup sequences, it does not appear to explain the extreme branch length differences over the first 2000 most varied positions in the OV alignment.

Fit of Data to a GTR+I+Γ Substitution Model

The effect of removal of the most varied sites on the fit of the aligned data to a GTR+I+Γ substitution model was investigated. Table 1 reports log-likelihood scores for 2 tree models (*Amborella* most basal; *Amborella*+*Trithuria*+Nymphaeaceae most basal) on A partitions generated by the script Sorter.pl. These scores were compared against the log-likelihood scores for data sets identical in length to the shortened A partitions that were jackknife resampled from the OV alignment. They were always significantly better than the scores for the randomly resampled data, indicating that the sites removed by OV noise reduction significantly contribute to the poor fit between the evolutionary models and the aligned sequence data.

Assuming the same evolution models as examined in Table 1, MISFITS and Tree-Puzzle were used to the identify site patterns whose relative frequencies are over- and underrepresented in the OV alignment. Figure 4 plots the position of unexpected site patterns in the ungapped OV alignment. The height of each bar in the histogram indicates the number of consecutive sites at which unexpected site patterns occur. The most varied end of the OV alignment is identified as containing many site patterns that contribute to the poor fit of the GTR+I+Γ substitution model.

Tree Building

Phylogenetic trees were built from the OV alignment for the different length A partitions generated by the Sorter.pl script. This was done both for a CAT+GTR+Γ+covext model and for a GTR+I+Γ model. The former was found under cross-validation to be optimal for: (i) the full-length OV alignment, (ii) the alignment of the first and the second codon positions, and (iii) the alignment of the most conserved 38 553 positions in the OV alignment. The optimal tree reconstructed with a CAT+GTR+Γ+covext model on the A partition at the GNB stopping point is

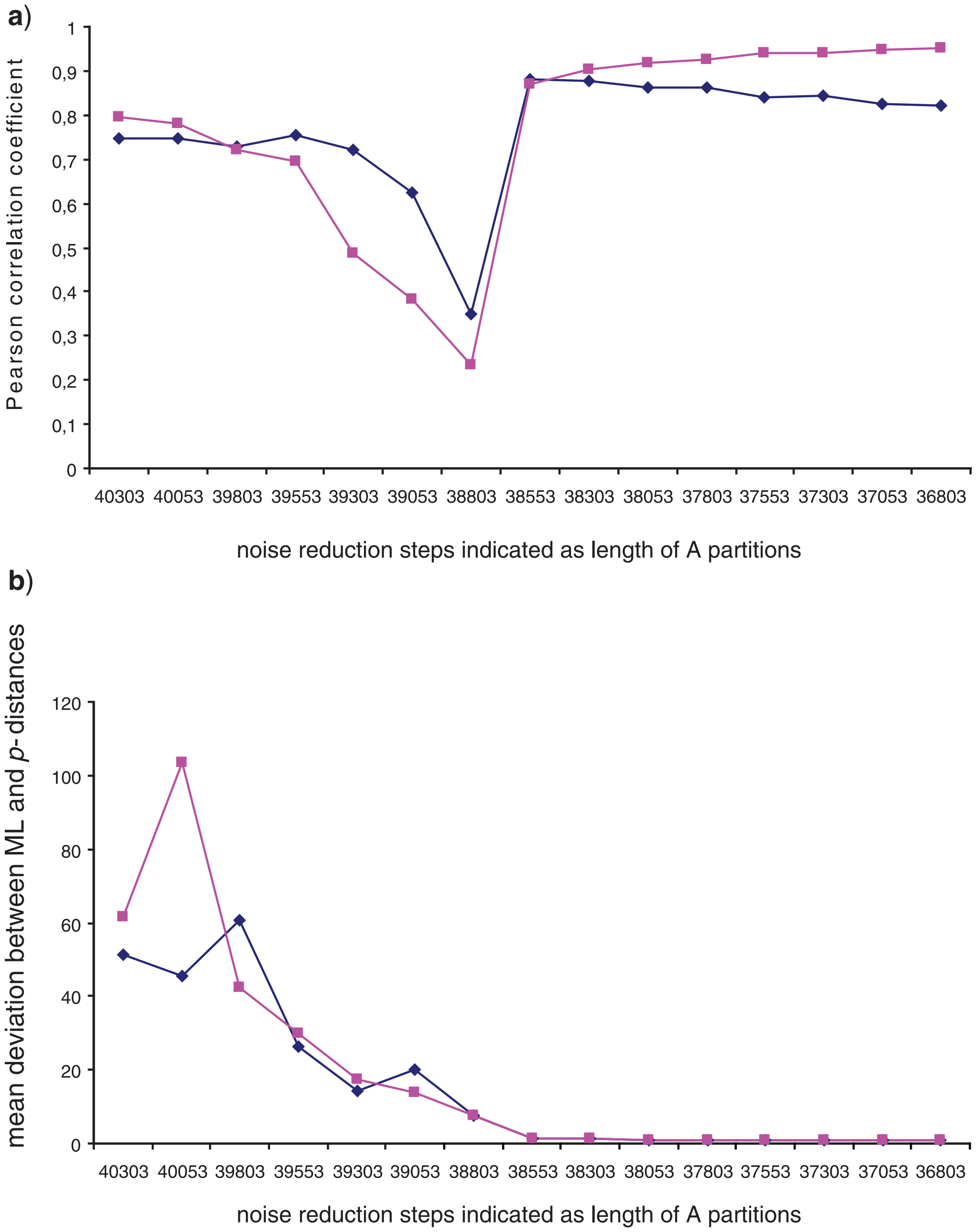

FIGURE 1. a) Plot showing results of the correlation analyses. The blue line indicates Pearson correlation coefficient values (r) obtained for pair-wise comparisons of ML-distances calculated from the A and B partitions whose combined length was 40 553 gapped positions in the OV alignment. The pink line indicates r values obtained for pair-wise comparisons of p-distances and ML-distances calculated for B partitions, discarded at each shortening step. At the 8th shortening step, when the A partition is 38 553 gapped positions in length, it passes both correlation tests (Goremykin et al. 2010). b) Plot showing mean deviation between ML- and p-distances calculated for B partitions at each shortening step. In calculating ML-distances, the best-fitting ML model for each partition length was first determined under an AIC using ModelTest (Posada and Crandall 1998). The pink line indicates results from analyses using a Neighbor-Joining tree to fit ML model parameters. The blue line indicates results obtained when an ML tree is used to fit substitution model parameters. This ML tree was computed using settings of the best-fitted model determined by the standard ModelTest procedure employing AIC.

shown in Figure 5. This tree indicates the same relationships among basal angiosperms as does the GTR+I+Γ tree reconstructed on the A partition at the GNB stopping point. Both reconstructions identify a lineage comprising *Amborella*+*Trithuria*+ Nymphaeaceae as most basal in the angiosperm radiation. Figure 6a indicates relationships inferred when a CAT+GTR+Γ+covext model is used to analyze the full-length (40 553 site) concatenated data set. With this data set, *Amborella* is inferred to be the most basal lineage in the radiation of angiosperms. Trees built from the alignment of first and second codon positions using

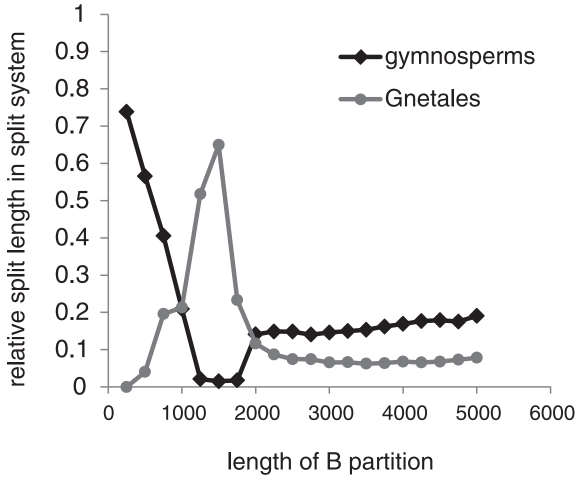

FIGURE 2. Plot showing the relative size of NNET splits separating: (i) angiosperms from gymnosperms and (ii) Gnetales from other taxa. The NNET splits were calculated from the optimal distances estimated for each B partition formed at the most varied end of the OV alignment.

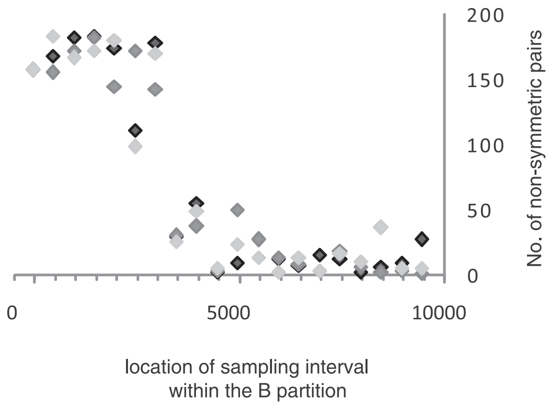

FIGURE 3. The number of pairwise distances (645 comparisons) failing a matched-pairs test of symmetry at $P < 0.00005$ was determined for equal length, nonoverlapping intervals at the most varied end of the OV alignment. For these estimates, we analyzed only ungapped sites (360 positions: 3 replicates per estimate) randomly sampled without replacement from 500-bp nonoverlapping gapped partitions at the most varied end of the OV alignment ("C" partitions in Goremykin et al. 2010).

the best-fitting CAT+GTR+Γ+covext model (Fig. 6b) show *Amborella*+*Trithuria*+Nymphaeaceae as the most basal lineage. Substitution models rejected in cross-validation supported the tree with the most basal branch subtending *Trithuria*+Nymphaeaceae (e.g., GTR+Γ model, Fig. 6c) based on the first and second position data set.

The support for relationships among basal angiosperms under a CAT+GTR+Γ+covext covarion model was also investigated after each shortening step of 250 positions in the alignment of all codon positions. The results are shown in Figure 7. These indicate that (i) support for *Amborella* joining with the outgroup occurs only when the most varied positions of the alignment are included, (ii) the grouping of *Amborella*+*Trithuria*+Nymphaeaceae is strongly favored as the most basal lineage after removal of 1750 sites and remains supported until 2500 sites are removed, and (iii) a basal grouping of *Amborella*+*Trithuria*+Nymphaeaceae+*Illicium* becomes favored after removal of 2750 sites. Note that under the CAT+GTR+Γ+covext model, support for *Amborella*+*Trithuria*+Nymphaeaceae as a most basal clade is realized prior to the GNB stopping point, which might indicate a better fit of this substitution model to the data.

DISCUSSION

Our findings reported here, and those in recent analyses of other seed plants (Zhong et al. 2011), reemphasize the importance of considering the fit of time-reversible models to the fast-evolving sites in sequence alignments (Sullivan et al. 1995). We show that site sorting can facilitate studies of the substitution properties of concatenated gene alignments and help to identify site patterns relevant to substitution model misspecification and potential tree-building artifacts. The sites providing most support for the *Amborella* most basal hypothesis are characterized by poor fit between model and data and by evolutionary properties that induce extreme topological distortion in reconstructed trees. The GNB stopping criterion removes many of these sites (38% of the removed sites did not fit an *Amborella* basal+GTR+I+Γ model; 39% of the removed sites did not fit an *Amborella*+*Trithuria*+Nymphaeaceae basal+GTR+I+Γ model).

In the present study, when sites causing topological distortion were removed, reconstruction under the optimal CAT model and GTR+I+Γ model favors a tree indicating *Amborella*+*Trithuria*+Nymphaeaceae as the most basal hypothesis. Although compositional heterogeneity will contribute to topological distortion when time-reversible Markov models are used in analysis of the data, our heterotachy and matched-pairs test of symmetry plots suggest that compositional heterogeneity is alone insufficient to explain the different topologies obtained during tree building with different A partitions. In general, the impact of compositional heterogeneity needs to be evaluated in the context of the extent of divergence between sequences exhibiting this heterogeneity (Jermiin et al. 2004) and the spatial pattern of sites free to vary in the sequences (Lockhart et al. 2006).

We propose that our analyses and observations provide a basis for understanding the discrepancy among recent findings from phylogenetic analyses of cpDNA and mtDNA concerning the rooting of the angiosperm phylogeny. Our reconstructed phylogeny (Fig. 5) obtained after exclusion of a large number of model-violating sites is consistent with that recently obtained in analyses of nuclear EST amino acid sequences that also implemented a CAT model. In this case, although *Trithuria* was not available for study,

TABLE 1. Data-model fit after removal of 500, 1000, 1500, and 2000 sites

Tree model	*Amborella*+Nymphaeaceae+*Ttithuria* most basal			
Number of sites retained	40053	39553	39053	38553
Mean log-likelihood values from jackknife samples	−332368.96	−328151.34	−323995.76	−319864.78
Log-likelihood values of shortened OV alignment	−321728.05	−307453.85	−294354.23	−282625.00
SD of jackknife samples	181.75	259.90	318.15	365.44
z-score	58.55	79.64	93.17	101.91
Tree model	***Amborella* most basal**			
Number of sites retained	40053	39553	39053	38553
Mean log-likelihood values from jackknife samples	−332328.03	−328110.72	−323955.46	−319825.17
Log-likelihood values of shortened OV alignment	−321708.11	−307439.72	−294347.58	−282630.57
SD of jackknife samples	181.64	260.30	318.54	365.80
z-score	58.47	79.41	92.95	101.68

The *z*-score is the difference between the mean log-likelihood value from jackknife samples and the log-likelihood value of the shortened OV alignment (equivalent length A partition). This difference is expressed in terms of number of standard deviations (SD) calculated for the jackknife samples. The improvement in data-model fit obtained by excluding sites at the most varied end of the OV alignment was always significant at $P < 0.001$ (no score for any jacknife sample was better than the score generated by noise reduction).

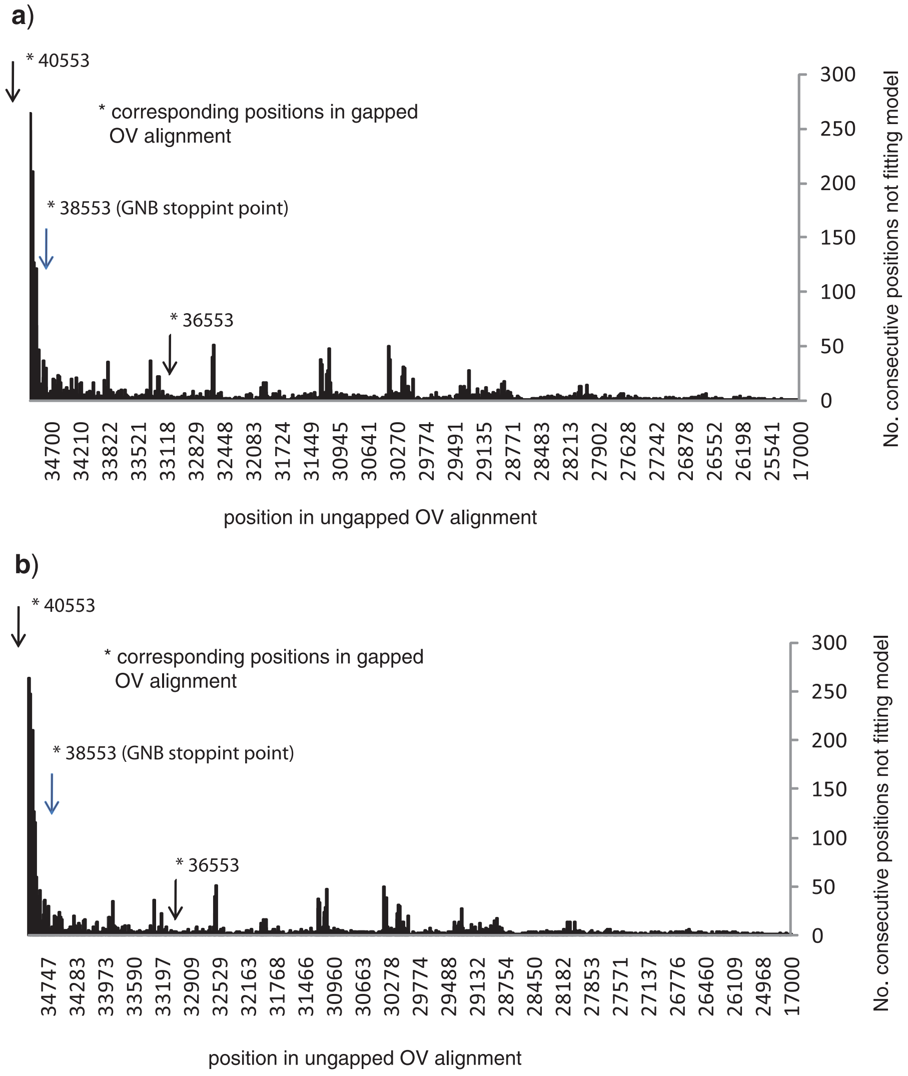

FIGURE 4. a) Histogram showing positions of sites in the OV alignment that contain site patterns unexpected under a GTR+I+Γ substitution model and *Amborella*+*Trithuria*+Nymphaeaceae hypothesis. b) Histogram showing positions of sites in the OV alignment that contain site patterns unexpected under a GTR+I+Γ substitution model and *Amborella* most basal hypothesis. A feature of both graphs is that relatively few sites fit either model at the most varied end of the OV alignment. Both ungapped positions and gapped positions (∗) have been indicated on the figure.

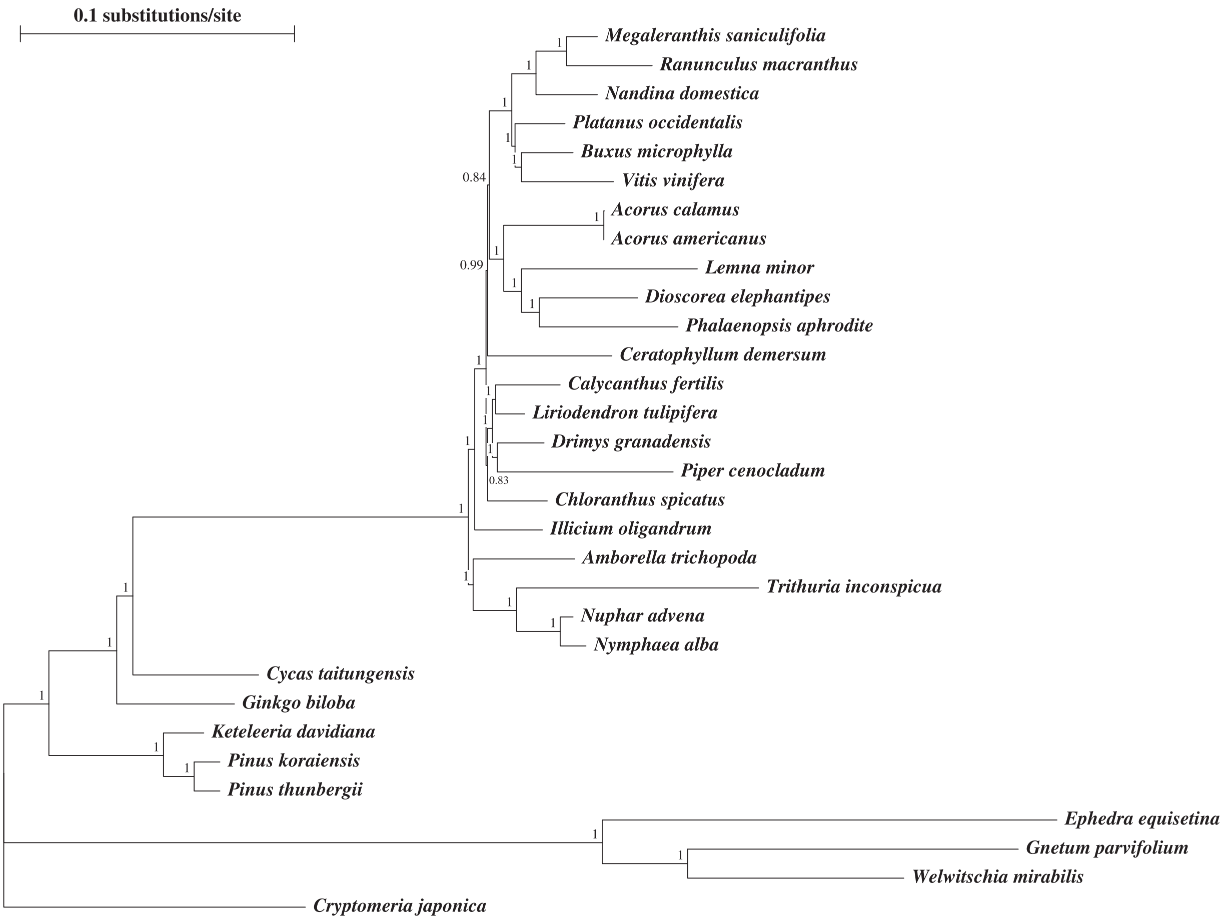

FIGURE 5. Tree reconstructed from Bayesian analysis and best-fitting substitution model (CAT+GTR+Γ+covext model) for the conserved A partition (38 553 sites) identified by the GNB criterion.

Amborella and *Nuphar* were inferred to be sister taxa (Finet et al. 2010). Our reconstruction is also congruent with recent analyses of 4 slowly evolving mitochondrial genes (Qiu et al. 2010).

Our phylogenetic reconstruction differs from that obtained in a recent and well-sampled ML-based phylogenetic analyses for 17 concatenated nuclear, mitochondrial, and chloroplast genes (Soltis et al. 2011). This study reported *Amborella* as most basal. Reanalyzing these data with a GTR+I+Γ model and RaxML, we were unable to confirm this finding. Rather, we inferred a phylogenetic tree wherein a clade comprising *Amborella*, *Trithuria*, and Nymphaeaceae received 94% nonparametric bootstrap support (results not shown). Whether this result indicates a shortcoming of the heuristic search with RaxML or a more accurate reconstruction of angiosperm phylogeny from this joint data matrix requires further investigation.

We conclude that analyses of available sequence data do not support the earliest angiosperms being woody and terrestrial. Evidence from phylogenetic analyses of concatenated chloroplast genes appears equally consistent with some of the earliest species being herbaceous and aquatic. Further tests of this hypothesis are needed. We suggest that our analytical protocol provides a valuable approach, and one that is potentially useful for other questions currently being investigated with phylogenomic data sets.

SUPPLEMENTARY MATERIAL

Data files and/or other supplementary information related to this paper have been deposited at Dryad under doi:10.5061/dryad.vs49s (www.datadryad.org).

FUNDING

PJL thanks the Allan Wilson Centre and New Zealand Royal Society (James Cook Fellowsship scheme and Marsden Fund) for financial support. WM also thanks the Deutsche Foschungsgemeinschaft (DFG) for funding.

ACKNOWLEDGEMENTS

We thank Dr. Lars Jermiin and the reviewers for their constructive and helpful suggestions.

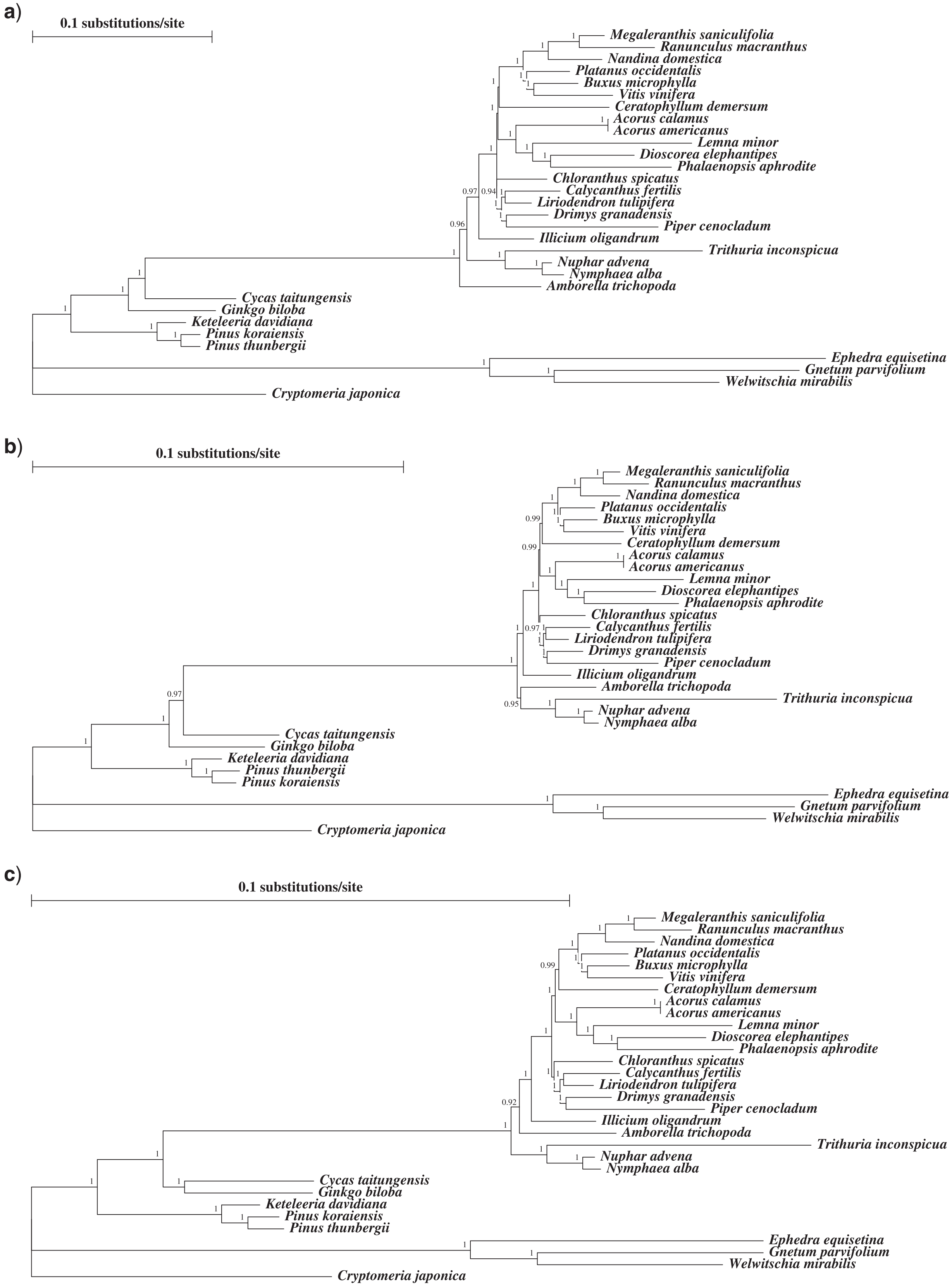

FIGURE 6. a) Tree reconstructed from Bayesian analysis and best-fitting substitution model (CAT+GTR+Γ+covext model) for the full-length (40 553 site) concatenated data set. b) Tree built from the alignment of the first and the second codon positions employing best-fitting CAT+GTR+Γ+covext model. c) Tree built from the alignment of the first and the second codon positions employing the GTR+Γ model.

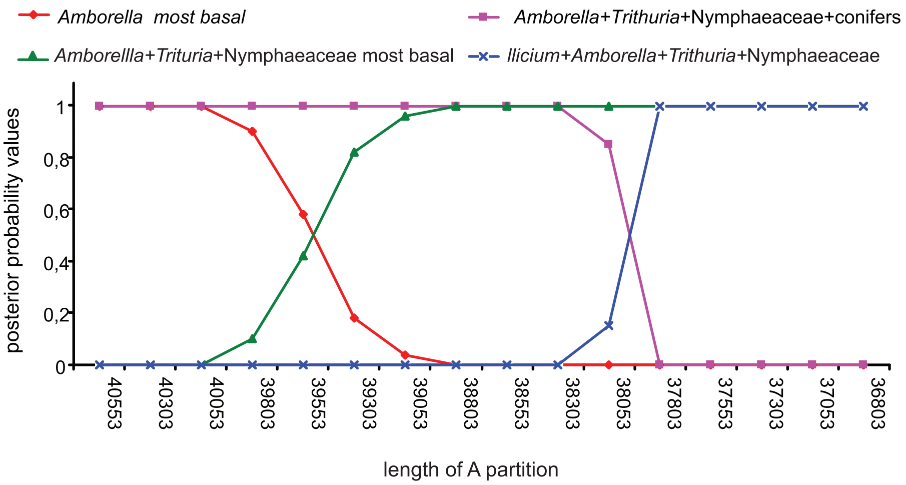

FIGURE 7. Posterior probability support for alternative hypotheses of relationship as sites are removed from the most varied end of the OV alignment computed under the best-fitting substitution model (CAT+GTR+Γ+covext). Similar inferences were obtained with taxon subsets that excluded the most compositionally heterogeneous sequences.

REFERENCES

Ababneh F., Jermiin L.S., Ma C., Robinson J. 2006. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. Bioinformatics 22:1225–1231.

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

Atherton R.A., McComish B.J., Shepherd L.D., Berry L.A., Albert N.W., Lockhart P.J. 2010. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods. 6:22.

Barkman T.J., Chenery G., McNeal J.R., Lyons-Weiler J., Ellisens W.J., Moore G., Wolfe A.D., dePamphilis C.W. 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.

Braukmann T.W., Kuzmina M., Stefanovié S. 2009. Loss of all plastid *ndh* genes in Gnetales and conifers: extent and evolutionary significance for the seed plant phylogeny. Curr. Genet. 55:323–337.

Brinkmann H., Philippe H. 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16:817–825.

Bruno W.J., Halpern A.L. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564–566.

Bryant D., Moulton V. (2004). Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21:255–265.

Burleigh J.G., Mathews S. 2004. Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. Am. J. Bot. 91:1599–1613.

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

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Drummond A.J., Ashton B., Buxton S., Cheung M., Cooper A., Heled J., Kearse M., Moir R., Stones-Havas S., Sturrock S., Thierer T., Wilson A. 2010. Geneious v5.1, Available from: URL http://www.geneious.com.

Edgar R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

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

Finet C., Timme R.E., Delwiche C.F., Marletaz F. 2010. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Curr. Biol. 20:2217–2222.

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

Goremykin V.V., Hirsch-Ernst K.I., Woelfl S., Hellwig F.H. 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., Hirsch-Ernst K.I., Woelfl S., Hellwig F. 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., Nikiforova S.V., Bininda-Emonds O.R.P. 2010. Automated removal of noisy data in phylogenomic analyses. J. Mol. Evol. 71:319–331.

Graham S.W., Iles W.J.D. 2009. Different gymnosperm outgroups have (mostly) congruent signal regarding the root of flowering plant phylogeny. Am. J. Bot. 96:216–227.

Gruenheit N., Lockhart P.J., Steel M.A., Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites Mol. Biol. Evol. 25:1512–1520.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hansmann S., Martin W.T. 2000. Phylogeny of 33 ribosomal and six other proteins encoded in an ancient gene cluster that is conserved across prokaryotic genomes: influence of excluding poorly alignable sites from analysis. Int. J. Syst. Evol. Microbiol. 50: 1655–1663.

Hirt R.P., Logsdon J.M., Healy B., Dorey M.W., Doolittle W.F., Embley T.M. 1999. Microsporidia are related to Fungi: evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Natl. Acad. Sci. U.S.A. 96:580–585.

Ho J.W.K., Adams C.E., Lew J.B., Matthews T.J., Ng C.C., Shahabi-Sirjani A., Tan L.H., Zhao Y., Easteal S., Wilson S.R., Jermiin L.S. 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. Bioinformatics 22:2162–2163.

Huson D.H., Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23:254–267.

Jansen R.K., Cai Z., Raubeson L.A., Daniell H., dePamphilis C.W., Leebens-Mack J., Mueller K.F., Guisinger-Bellian M., Haberle R.C., Hansen A.K., Chumley T.W., Lee S.-B., Peery R., McNeal J.R., Kuehl J.V., Boore J.L. 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc. Natl. Acad. Sci. U.S.A. 104:19369–19374.

Jermiin L., Ho S. Y.W., Ababneh F., Robinson J., Larkum A.W.D. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.

Jiao Y., Wickett N.J., Ayyampalayam S., Chanderbali S., Landherr L., Ralph P.E., Tomsho L.P., Hu Y., Liang H., Soltis P.S., Soltis D.E., Clifton S.W., Schlarbaum S.E., Schuster S.C., Ma H., Leebens-Mack J., dePamphilis C.W. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473:97–100.

Lartillot N., Philippe H. 2004. A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.

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

Lockhart P. J., Beanland T. J., Howe C. J., Larkum A.W.D. 1992. Sequence of *Prochloron didemni* atpBE and the inference of chloroplast origin. Proc. Natl. Acad. Sci. U.S.A. 89:2742–2746.

Lockhart P.J., Larkum A.W.D Steel M.A., Waddell P.J., Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis.Proc. Natl. Acad. Sci. U.S.A. 93:1930–1934.

Lockhart P.J., Novis P., Milligan B.G., Riden J., Rambaut A., Larkum A.W.D. 2006. Heterotachy and tree building: a case study with plastids and Eubacteria Mol. Biol. Evol. 23:40–45.

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

Lopez P., Forterre P., Philippe H. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49:496–508.

Martin W.T., Deusch O., Stawski N., Gruenheit N., Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10: 203–205.

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

Nguyen M.A.T., Klaere S., von Haeseler A. 2011. MISFITS: evaluating the goodness of fit between a phylogenetic model and an alignment. Mol. Biol. Evol. 28:143–152.

Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Woerheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Pisani D. 2004. Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. Syst. Biol. 53:978–989.

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

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. The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402:404–407.

Qiu Y.-L., Wang B., Xue J.-Y., Hendry T.A., Li R.-Q., Brown J. W., Liu Y., Hudson G.T., Chen Z.-D. 2010. Angiosperm phylogeny inferred from sequences of four mitochondrial genes. J. Syst. Evol. 48: 391–425.

Rambaut A. 2002. Se-Al. Sequence Alignment Editor v2.0a11. Available from: URL http://evolve.zoo.ox.ac.uk.

Rodriguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B.F., Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56:389–399.

Ruiz-Trillo I., Riutort M., Littlewood D.T., Herniou E.A., Baguna J. 1999. Acoel flatworms: earliest extant bilaterian metazoans, not members of Platyhelminthes. Science 283:1919–1923.

Saarela J.M., Rai H.S., Doyle J.A., Endress P.K., Mathews S., Marchant A.D., Briggs B.G., Graham S.W. 2007. Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature 446:5–8.

Schmidt H.A., Strimmer K., Vingron M., von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18: 502–504.

Shavit L., Penny D., Hendy M.D., Holland B.R. 2007. The problem of rooting rapid radiations. Mol. Biol. Evol. 24:2400–2411.

Soltis D.E., Smith S.A., Cellinese N., Wurdack K.J., Tank D.C., Brockington S.F., Refulio-Rodriguez N.F., Walker J.B., Moore M.J., Carlsward B.S., Bell C.D., Latvis M., Crawley S., Black C., Diouf D., Xi Z., Rushworth C.A., Gitzendanner M.A., Sytsma K.J., Qiu Y.L., Hilu K.W., Davis C.C., Sanderson M.J., Beaman R.S., Olmstead R.G., Judd W.S., Donoghue M.J., Soltis P.S. 2011. Angiosperm phylogeny: 17 genes, 640 taxa. Am. J. Bot. 98:704–730.

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

Sperling E.A., Peterson K.J., Pisani D. 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of eumetazoa. Mol. Biol. Evol. 26:2261–2274.

Stamatakis A., Ludwig T., Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.

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

Sullivan J., Holsinger K.E., Simon C. 1995. Among-site variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. Mol. Biol. Evol. 12:988–1001.

Sun G., Dilcher D.L., Wang H., Chen Z. 2011. A eudicot from the Early Cretaceous of China. Nature 471:625–628.

Swofford D.L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland (MA): Sinauer Associates.

Tamura K., Peterson D., Peterson N., Stecher G., Nei M., Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

Tuffley C., Steel M.A. 1998. Modelling the covarion hypothesis of nucleotide substitution. Math. BioSci. 147:63–91.

Wakasugi T., Tsudzukit J., Itot S., Nakashimat K., Tsudzuki T. 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.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Wu C.-S., Wang Y.-N., Liu S.-M., Chaw S.-M. 2007. Chloroplast genome (cpDNA) of *Cycas taitungensis* and 56 cp protein-coding genes of *Gnetum parvifolium*: insights into cpDNA evolution and phylogeny of extant seed plants. Mol. Biol. Evol. 24: 1366–1379.

Yang Z. 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.

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

Zerbino D.R., Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18: 821–829.

Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3: 1340–1348.