The Root of Flowering Plants and Total Evidence

V. V. Goremykin1,∗, S. V. Nikiforova1, D. Cavalieri1, M. Pindo1, and Peter Lockhart2,3

1FEM Research and Innovation Center, Via E. Mach 1, 38010 San Michele all’Adige (TN), Italy
2Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand, and
3Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

∗Correspondence to be sent to: FEM Research and Innovation Center, Via E. Mach 1, 38010 San Michele all’Adige (TN), Italy.
E-mail: Vadim.Goremykin@fmach.it

Received 13 October 2014; reviews returned 9 March 2015; accepted 5 May 2015

Associate Editor: Edward Susko

Abstract.—Support for Amborella as the sole survivor of an evolutionary lineage that is sister to all other angiosperms comes from positions in DNA multiple-sequence alignments that have a poor fit to time-reversible substitution models. These sites exhibit significant levels of homoplasy, compositional heterogeneity, and strong heterotachy. We report phylogenetic analyses with observed, randomized, and simulated data which show there is little or no expectation that these sites provide useful information for understanding relationships among basal angiosperms. Their inclusion in phylogenetic analyses leads to a long-branch attraction artifact that favors Amborella as sister to other angiosperms in reconstructed phylogenies. Using parametric simulations, we show that sites in chloroplast sequences that exhibit less homoplasy between angiosperms and gymnosperms provide more reliable information for inferring basal angiosperm relationships. We confirm our earlier findings that the basal angiosperm Amborella is most closely related to aquatic herbs. Our current and previously reported (Goremykin et al. 2013) analyses highlight an essential aspect of the total evidence approach to phylogenetic inference. They suggest that data partitioning aimed at identifying components of the data that better fit evolutionary models is a more reliable approach to phylogeny reconstruction at deep taxonomic levels. [Amborella, angiosperm origins; model fit; parametric simulations; systematic error; total evidence approach.]

Evolutionary trees which show Amborella as sister to all other angiosperms have been commonly reported in systematic molecular investigations, but typically only when alignment positions with high levels of character state variation are included in data matrices (Zanis et al. 2002; Stefanović et al. 2004; Leebens-Mack et al. 2005; Jansen et al. 2007; Moore et al. 2010; Solits et al. 2011; Drew et al. 2014). A close phylogenetic relationship between Amborella and aquatic herbs is also a relationship that has been repeatedly recovered by researchers, in this case typically when sites showing less character state variation are analyzed (Barkman et al. 2008; Solits et al. 2008; Zanis et al. 2002; Chang et al. 2005; Leebens-Mack et al. 2005; Qiu et al. 2005, 2006, 2010; Bausher et al. 2006; Jansen et al. 2006; Moore et al. 2007; Wu et al. 2007; Mardanov et al. 2008; Graham and Iles 2009; Finet et al. 2010; Goremykin et al. 2009, 2013; Jiao et al. 2011; Wodniok et al. 2011; Laurin-Lemay et al. 2012; Xi et al. 2014). Goremykin et al. (2013) raised concerns that the Amborella most basal placement commonly reported is the result of a phylogenetic artifact due to systematic error and a poor fit between time-reversible substitution models and sequence data. The same article showed that with chloroplast sequences this poor fit was mainly due to sequence positions in multiple-sequence alignments that exhibit high levels of character state variation among angiosperms and gymnosperms. More recently, Xi et al. (2014) have suggested that a similar problem affects nuclear sequences. These authors used parametric simulations to show that inference of the Amborella most basal root placement is unreliable with nuclear sequences and that optimal phylogenetic reconstruction for conserved chloroplast and nuclear genes recover Amborella adjacent to aquatic herbs, including Trithuria and Nymphaeales. Xi et al.’s (2014) work is significant as it extends Goremykin et al.’s (2013) finding of a poor data-model fit at fast-evolving sites in chloroplast sequences to the fast-evolving sites of nuclear sequences, and shows that these sites will also cause an error in phylogenetic analyses of angiosperm origins. Xi’s et al.’s (2014) analyses did not explain why these fast-evolving sites cause long-branch attraction (LBA) with chloroplast and nuclear sequences. We address this issue here.

Recently, Drew et al. (2014) emphasized that the fast-evolving sites needed for an Amborella most basal hypothesis are informative and that these should be included in phylogenetic analyses. One reason Drew et al. (2014) discounted the findings of Goremykin et al. (2013) concerns objection to the observed variability (OV)-based sorting protocol used by the authors to group site patterns and to study their evolutionary properties. This is the same sorting protocol used recently by Xi et al. (2014)—who also employed the TIGER site sorting protocol of Cummins and McInerney (2011)—to obtain further results that support the findings of Goremykin et al. (2013).
Recognition of the ways in which model mis-specification misleads phylogenetic inference has driven methodology development (Lockhart et al. 1994; Ané et al. 2005; Waddell 2005; Ababneh et al. 2006; Lartillot and Philippe 2008; Jayaswal et al. 2014). Thus, it is instructive to further examine why LBA has missed so many researchers regarding basal relationships of flowering plants. Central to this issue is adoption of the “total evidence” (TE) approach for phylogenetic analysis. While the principle of TE (introduced by Kluge 1989, following Carnap 1950), merely means using all the evidence available, one particular interpretation of TE advocates combining all available data (taxa and characters) into a single data matrix under the assumption that this will lead to better results. The most extreme view of this—adopted by a number of plant systematists (Chase et al. 1995; Källersjö et al. 1999; Savolainen et al. 2000; Hilu et al. 2003; Solé et al. 2004)—has perhaps been most clearly stated by Savolainen et al. (2000) “Homoplasy is evidence, and the more evidence that is available, the more accurate is the resulting tree.” This perspective argues that discarding extremely variable characters—as done by Goremykin et al. (2013)—can result in loss of valuable phylogenetic signal. This view further holds that “the utility of characters for phylogenetic analysis cannot be determined a priori on the basis of character variability” (Drew et al. 2014).

A point perhaps not appreciated when these criticisms were directed at Goremykin et al. (2013) was that it was not character variability, but rather substitution model mis-specification that was used as the criterion for removing sites that exhibited extreme character state variation. This was done because there are well-documented concerns about the reliability of tree building when the evolutionary properties of DNA sequences are not well described by the assumptions of the substitution model. These concerns began to be raised at least 30 years ago (e.g., Lanave et al. 1984; Penny et al. 1992; Hasegawa and Hashimoto 1993) and they continue to this day (e.g., Cooper 2014). One of the first empirical and theoretical examples of the importance of this issue for maximum-likelihood (ML) inference was demonstrated by Lockhart et al. (1996). These authors showed that when the substitution model assumed all sites were variable, but where in fact, some invariable, ML would fail to correctly estimate branch length differences caused by lineage-specific rate variation and could fail to recover the correct topology. Many others have since also demonstrated ML to be inconsistent where substitution models are mis-specified in one way or another (e.g., Chang 1996; Kolačzkowski and Thornton 2004; Schwarz et al. 2004; Spencer et al. 2005; Thornton and Kolačzkowski 2005). Association of model mis-specification and systematic error with fast-evolving sequence positions is now well established in the phylogenetic literature (Steel et al. 1993; Lockhart et al. 1994; Brinkmann and Philippe 1999; Hirt et al. 1999; Lopez et al. 1999; Ruiz-Trillo et al. 1999; Hansmann and Martin 2000; Pisani 2004; Delsuc et al. 2005; Philippe et al. 2005; Jeffroy et al. 2006; Burleigh and Mathews 2007; Rodriguez-Ezepeleta et al. 2007; Sperling et al. 2009; Cummins and McInerney 2011; Pisani et al. 2012, etc.).

Systematic error is an issue for reconstructing basal angiosperm relationships because site saturation at the fastest evolving sites of chloroplast coding gene sequences occurs within the taxonomic range of the taxa studied. This feature of the data has long been recognized (Goremykin et al. 1996, 2003; Chaw et al. 2000, 2004; Qi et al. 2006, 2010). Goremykin et al. (2013) showed that the fastest evolving sites in these data are characterized by both compositional heterogeneity and lineage-specific rate variation (heterotachy) that contributes to extreme branch-length differences between in-group and out-group taxa.

Furthermore, the observation that compositional heterogeneity and heterotachy are most strongly exhibited at sites with the greatest character state variation (Goremykin et al. 2013) means that these sites cannot be easily partitioned and modeled as a time-reversible substitution process. This problem has been previously observed and discussed (e.g., Lockhart and Steel 2005; Lockhart et al. 2006; Jayaswal et al. 2014).

Here we attempt to bring resolution to the controversy over Amborella by demonstrating that the evolutionary properties of the data studied by Goremykin et al. (2013) are similar to the evolutionary properties of the data studied by Drew et al. (2014). We explain why the results of Drew et al. (2014) differ from those of Goremykin et al. (2013), and also explain why Amborella is falsely drawn to the root of the angiosperm phylogeny by LBA.

**Materials and Methods**

**Data Sets**

Alignment S1: the alignment of Drew et al. (2014).—We analyzed the aligned data matrix of Drew et al. (2014), a 78-gene, 236-taxon aligned data matrix provided as Supplementary File S1 to that paper.

Alignment 1: a 36 OTU reduced data set from Drew et al. (2014).—To build trees based on site-heterogeneous models within a reasonable time frame, we limited sampling among eudicots and monocots. We constructed a data set of 36 taxa which represented all basal angiosperm and magnoliid lineages present in the S1 alignment of Drew et al. (2014). Angiosperm taxa excluded from alignment S1 were species which have little impact on the relationship among basal angiosperms and their relationship to the gymnosperm outgroup (Fig. 2a). To demonstrate that the findings in Goremykin et al. (2013) were not affected by the presence of Gnetales, as speculated by Drew et al. (2014), we excluded Gnetales from alignment 1, leaving all other gymnosperms. We deleted gap-only sites from the taxon-wise reduced alignment. The resulting 58,554
Each A partition was used as in previous studies (Goremykin et al. 2010, 2013). The partitioning was performed conserving (B partitions) subsets using the sorter.pl script into bipartitions: conserved (A partitions) and less conserved (covext) models. Five chains were run under each model/data combination for 2000 cycles. We discarded 500 cycles as burn-in for all chains to reach maximal ML values, and sampled every cycle thereafter to build trees.

Alignment 2: first and second codon position data set from Goremykin et al. (2013).—To investigate the suggestion made by Drew et al. (2014) that analyses of the first and the second positions in the in-frame alignment reported in Goremykin et al. (2013) support the hypothesis of Amborella being the sole representative of a basal-most angiosperm lineage, we extracted the first and the second positions from the in-frame alignment (Goremykin et al. 2013), and created a separate file containing these sites (21,116 pos. long alignment 2, Supplementary File 2 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f). Trees were built for the first 24 B partitions of the resulting two sorted alignments using RAxML and the GTR+G4 site-homogeneous substitution model. For every optimal RAxML tree recovered from a given B partition, we recorded: (i) the length of the branches subtending the angiosperm outgroup and (ii) the proportion of the branch subtending the out-group with respect to total tree length (Fig. 1).

To explain discrepancies in the findings of Drew et al. (2014) and Goremykin et al. (2013), we compared sorted alignments obtained when OV scores were based on different subsets of taxa. This comparison showed how taxon sampling affects OV scores and impacts on the effectiveness of OV sorting in identifying saturated sites between the in-group and the out-group.

For this analysis, we took the unedited 236-taxon SI alignment from Drew et al. (2014) and subjected it to the sorting procedure, once using OV scores estimated for the 36 taxa included in alignment 1, and once using OV scores obtained with the complete (236) OTU set. We analyzed B partitions sampled from the SI alignment sorted with 36-taxon OV scores (provided as Supplementary File 5 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f) and from the SI alignment sorted with 236-taxon OV scores (provided as Supplementary File 6 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f). Trees were built for the first 24 B partitions of the resulting two sorted alignments using RAxML and the GTR+G4 site-homogeneous substitution model. For every optimal RAxML tree recovered from a given B partition, we recorded: (i) the length of the branches subtending the angiosperm outgroup and (ii) the proportion of the branch subtending the out-group with respect to total tree length (Fig. 1).

The impact of taxon (monocot and eudicot) sampling on angiosperm root placement was also assessed in analyses of the 36-taxon (I) and 236-taxon (SI) alignments. 36-taxon OV scores were used to order sequence positions in both alignments. Doing this allowed us to remove the same alignment positions at every shortening step, as well as to directly assess the added value of the denser monocot and eudicot sampling used by Drew et al. (2014) for inferring angiosperm root placement. We compared the bootstrap support values for basal angiosperm placements in ML trees for the entire sorted 36- and 236-taxon alignments and also for the first 24 A partitions of each ordered alignment, where each partition was decreased by a length of 250 sites (Fig. 2a).

Alignments 1 and 2 were used as a proxy for substitution rate, and the resulting sorted alignment (provided as Supplementary File 4 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f) was divided into bipartitions: conserved (A partitions) and less conserved (B partitions) subsets using the sorter.pl script (Goremykin et al. 2013). The partitioning was performed in intervals of \( n \times 250 \) (where \( n = 1, 2, 3 \ldots 19 \)) positions from the most varied end of the sorted alignment. Incremental shortening of partitions by 250 positions was used as in previous studies (Goremykin et al. 2000, 2013). Each A partition contained 58554-250×n sites and each B partition contained 250×n sites.

OV Scores Used to Order Data Sets from the Alignment of Drew et al. (2014)

Alignment 1 was sorted using OV scores as a proxy for substitution rate, and the resulting sorted alignment (provided as Supplementary File 4 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f) was divided into bipartitions: conserved (A partitions) and less conserved (B partitions) subsets using the sorter.pl script (Goremykin et al. 2013). The partitioning was performed in intervals of \( n \times 250 \) (where \( n = 1, 2, 3 \ldots 19 \)) positions from the most varied end of the sorted alignment. Incremental shortening of partitions by 250 positions was used as in previous studies (Goremykin et al. 2000, 2013). Each A partition contained 58554-250×n sites and each B partition contained 250×n sites.

Analyses of Alignments 2 and 3

We conducted Phylobayes analyses on alignment 2 specifying CAT+GTR+G and CAT+GTR+G+covext (wherein covext refers to a modification of Tuffley and Steel’s covarian model implemented in Phylobayes) models, and on alignment 3, specifying GTR+covext models. Five chains were run under each model/data combination until 2000 cycles were sampled. For every chain, we discarded the first 500 cycles as “burn-in” which was sufficient for all chains to reach maximal ML values, and sampled every cycle thereafter to build trees.
FIGURE 1. a) Lengths of branches connecting angiosperm subtrees to out-groups in RAxML trees built from variable B partitions of 236-taxon alignment S1 (Drew et al. 2014) obtained applying OV scores computed based on all 236 taxa (a line with diamonds) and applying OV scores estimated based on the 36-taxon subset (a line with squares). Values on the Y-axis indicate branch lengths [subst./site] and values on the X-axis indicate the length of corresponding B partitions. b) The same branch lengths measured as a percentage of the total length of trees estimated from B partitions obtained applying OV scores computed based on all 236 taxa (a line with diamonds) and obtained based on OV scores for the 36 taxa (a line with squares). Values on the Y-axis indicate percent values and values on the X-axis indicate length of corresponding B partitions.

distribution (CAT+GTR+G model) and comparing branch support for basal angiosperm relationships.

Four different unweighted trees were identified with the above analyses (Fig. 3, Supplementary Fig. 1 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f). These trees, which represent alternative hypotheses of relationships among basal angiosperms and gymnosperms, were used as model trees in the simulation analyses.

Phylogenetic Signal or Noise in the Fastest Evolving Sites?

To evaluate the usefulness of the fastest evolving sites in resolving relationships among basal angiosperms, we undertook three complementary approaches: (i) we conducted phylogenetic analyses of B partition data and equivalent length jackknife resampled A partition data; (ii) we conducted parametric simulations for the full as well as the partitioned data; and (iii) we compared relative site saturation in basal angiosperm sequences in B partitions, using an approach similar to that originally proposed in Steel et al. (1993, 1995).

Phylogenetic analyses of B partition data.—We examined support for outgroup placements in phylogenetic analyses of B partition sites from alignment I. Since trees built from the conserved A partitions under a CAT+GTR+G model begin to favor Amborella grouping
FIGURE 2. a) Bootstrap support values obtained for alternative basal-most angiosperm branches in unrooted RAxML trees built under a GTR+G4 model based on 36-taxon alignment 1 and its A partitions (lines with triangles represented on the right side of the legend) and based on 236-taxon S1 alignment (Drew et al. 2014) and its A partitions obtained using the same 36-taxon OV scores (lines with squares represented on the left side of the legend). Values on the Y-axis indicate bootstrap values and values on the X-axis indicate length of corresponding B partitions.

b) PP support obtained for alternative basal-most angiosperm branches and alternative sister groups to angiosperms in unrooted Phylobayes trees built under CAT+GTR+G4 model based on 36-taxon alignment 1 and its A partitions. Values on the Y-axis indicate the PP values and values on the X-axis indicate length of the corresponding B partitions.

c) PP support obtained for alternative basal-most angiosperm branches and alternative sister groups to angiosperms in unrooted Phylobayes trees built under CAT+GTR+continuous gamma model based on 36-taxon alignment 1 and its A partitions. Values on the Y-axis indicate the PP values and values on the X-axis indicate length of the corresponding B partitions.
with *Trithuria* and Nymphaeales (i.e., the ANT clade) with maximum PP support (Fig. 2c) from the sixth shortening (A₆) step on (i.e., after removal of at least 1500 sites with the highest OV scores), we built trees for the residual B₆ partition and compared these with trees built for 1500 pos. long data partitions randomly sampled from the full alignment.

We carried out Bayesian analyses under a CAT+GTR+G model 50 times for the B₆ partition of alignment 1, each time sampling 20,000 cycles and discarding the first 10,000 cycles to build trees. The very large cutoff of 10,000 cycles, reaching far into a plateau of ML scores, was chosen to highlight the absence of convergence of mature chains. Different attachments of the gymnosperm out-group registered in these experiments are shown in Figure 4 (Y-axis, experiment A). In parallel, we also analyzed 50 nonoverlapping jackknife replicates of the same size (1500 pos.) sampled from alignment 1 with the help of the Seqboot program from the Phylip v. 3.36 package. A single chain was run under CAT+GTR+G model for each jackknife replicate until 20,000 cycles were sampled. After discarding 500 cycles as "burn in," we built trees based on the remaining cycles. We compared various placements of the angiosperm root observed in trees built from the jackknife replicates (Fig. 4, Y-axis, experiment B) to the placements observed in analyses of the B₆ partition data.

We also repeated the above analyses for the 236-taxon data set of Drew et al. (2014) when the sequence positions were ordered by OV scores for the 36-taxon data set. This was done to investigate whether the increased taxon sampling in Drew et al. (2014) improved phylogenetic inference of basal angiosperm relationships from the B partition sites. For 50 replicates, we sampled 40,000 cycles in Phylobayes using the specification of the CAT+GTR+G model for the B₆ partition, and built trees after discarding the first 35,000 cycles. Again, this cutoff, reaching far into a plateau of ML scores, was chosen to highlight the absence of convergence of mature chains. The results of these experiments are presented in Figure 4 as experiment C. We also analyzed 50 jackknife replicates of the S₁ alignment which were of the same length as the B₆ partition (1500 pos.). These replicates were obtained using the Seqboot program. Twenty thousand cycles were sampled under a the CAT+GTR+G model, and we registered all alternative placements of the angiosperm root in trees built after discarding the first 5,000 cycles as

![Diagram](image-url)
Simulations assumed a CAT $+\Gamma_4$ root placements. We also conducted RAxML analyses, and used the same substitution model adopted by Drew et al. (GTR $+\Gamma_4$) to identify the best scoring ML tree among 100 ML trees built starting from 100 randomized MP trees for the B$_8$ partitions analyzed above (36- and 236-taxon data matrices).

**Parametric simulations.**—We used parametric simulations to evaluate the accuracy of the tree reconstruction method adopted by Drew et al. (2014) from full-length data (OV-sorted alignment 1) and its A and B partitions. For reconstructions, we used RAxML and assumed a GTR $+\Gamma_4$ model as used by Drew et al. (2014). Simulations assumed a CAT $+\Gamma_4$ model which was previously found in cross-validation experiments (Goremykin et al. 2013) to provide a good fit for a data set of concatenated chloroplast protein-coding genes. We simulated 36-taxon sequence alignments with Phylogenies based on OV-sorted alignment 1 and its A$_{16}$ and B$_{16}$ partitions that had been created using the sorter.pl script. The 16th shortening step was chosen because, with beginning this shortening step, a basal-most Amborella plus Trithuria plus Nymphaeae clade was consistently recovered with high (BP $\geq$ 80%) support under the GTR $+\Gamma_4$ model from A partitions (Fig. 2a), and we wished to determine if the angiosperm root placement could be confidently inferred from different data partitions which strongly supported contradictory root placements.

We ran chains, sampling for 2000 cycles, under the CAT $+\Gamma_4$ model for the A$_{16}$ and B$_{16}$ partitions and the full data set. For these simulations, four model tree topologies (Fig. 3, Supplementary Fig. 1 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f) previously estimated in unconstrained analyses of A partitions (Fig. 2) were enforced as alternative constraints. Five chains were run for each evolutionary model (tree plus substitution model combination). A total of 60 constrained chains were produced and saved (+option) at this experimental stage. For each of the 60 chains, we discarded the first 500 cycles as burn-in, which was found to be sufficient for all chains, and sampled 10 parametric replicates in intervals of 150 cycles with the help of the ppred program (distributed as a part of the Phylobayes package) run using posterior averages of the Bickel-in (Fig. 4, Y-axis, experiment D). A more detailed description of experiments summarized in Figure 4 is presented in Supplementary Table 1 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f.

To test whether recovery of the ANT clade from the conserved A$_{16}$ partitions could be attributed to biases of site deletion in the absence of model mis-specification, we OV-sorted each replicate which simulated the full alignment length and recorded the number of times RAxML recovered correct and spurious attachment of the angiosperm outgroup to the angiosperm subtree from the 54,554 pos. long A$_{16}$ partitions of the sorted replicates (shown in Supplementary Table 3 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f).

**Homoplasy and LBA among basal angiosperms.**—We also conducted RAxML analyses, and used the same substitution model adopted by Drew et al. (GTR $+\Gamma_4$) to identify the best scoring ML tree among 100 ML trees built starting from 100 randomized MP trees for the B$_8$ partitions analyzed above (36- and 236-taxon data matrices).

Detailed outcomes for these analyses are summarized in Supplementary Table 2 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f.

To test whether recovery of the ANT clade from the conserved A$_{16}$ partitions could be attributed to biases of site deletion in the absence of model mis-specification, we OV-sorted each replicate which simulated the full alignment length and recorded the number of times RAxML recovered correct and spurious attachment of the angiosperm outgroup to the angiosperm subtree from the 54,554 pos. long A$_{16}$ partitions of the sorted replicates (shown in Supplementary Table 3 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f). We also conducted RAxML analyses, and used the same substitution model adopted by Drew et al. (GTR $+\Gamma_4$) to identify the best scoring ML tree among 100 ML trees built starting from 100 randomized MP trees for the B$_8$ partitions analyzed above (36- and 236-taxon data matrices).

Homoplasy and LBA among basal angiosperms.—We examined the relative extent of site saturation for basal angiosperm (Nymphaea, Trithuria, and Amborella) sequences using a randomization approach based on the principles of the Steel et al. (1993, 1995) frequency-dependency test. The question was whether there were differences in site saturation among basal angiosperm sequences at their most varied sites as this might contribute to problems of LBA.

We analyzed B partition matrices of increasing length (for intervals between 250–2500 OV-sorted B partition positions) from alignment 1. We measured site saturation at all parsimony sites within these intervals, excluding sites at which indels were present. We chose to study parsimony sites (i.e., where there are at least 2 x 2 character states per site) because these sites contribute significantly to support for internal branches under Bayesian and likelihood inference methods. Our test assumed an Amborella most basal tree topology and we evaluated whether basal angiosperm sequences were random at these sites. To do this, we compared the support under a maximum parsimony criterion (i.e., a simple non-model-based counting method) for a fixed Amborella most basal tree topology before and after basal angiosperm sequences were individually randomized. Replicates were randomized in block using http://www.bioinformatics.org/sms2/shuffle_dna.html, last accessed June 26, 2015. Win-ParaP 4.10 (Swoford 2002) was used to calculate the parsimony scores and also the topologies of unconstrained parsimony trees.

**RESULTS**

Soltis et al. (2011) Data and the First and Second Codon Positions from In-frame Alignment of Goremykin et al. (2013)

All 10 Bayesian analyses of first and second codon positions of our previously published in-frame alignment (Goremykin et al. 2013) using the CAT+$\Gamma_4$, CAT+$\Gamma_4$+$\Gamma_4$, and CAT+$\Gamma_4$+$\Gamma_4$+$\Gamma_4$ models yielded trees that contained the basal-most ANT clade.

The clade was well supported both under the CAT+$\Gamma_4$+$\Gamma_4$+$\Gamma_4$ and CAT+$\Gamma_4$+$\Gamma_4$+$\Gamma_4$+$\Gamma_4$ models.

The question was whether there were differences in site saturation among basal angiosperm sequences at their most varied sites as this might contribute to problems of LBA.
model (1, 0.98, 0.99, 1, and 0.98 PP in five different analyses) and under the CAT+GTR+G model (1, 0.96, 0.98, 0.95, and 1 PP in five different analyses).

We compared results of phylogeny reconstruction based on Bayesian inference with our previously reported findings of strong ANT branch support from RAxML analysis (Goremykin et al. 2013) which were based on unmodified Soltis et al. (2011) alignment. All five separate Phylobayes runs on a taxon-reduced data set from Soltis et al. (2011) (alignment 3) using a site-homogeneous model (GTR+G) also recovered a well-supported (PP = 1) basal-most ANT clade. Under the site-heterogeneous CAT+GTR+G and CAT+GTR+G+covext models, the ANT clade was recovered with the same support in all 10 experiments (five per model) (Supplementary Fig. 2b available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f).

Fast-evolving Sites and OV Scores

Sorting using the 36-taxon OV scores was effective in identifying and concentrating in the B partition, site patterns that contribute to a large path length between in- and out-group, both in absolute terms and in comparison to the length of other branches. In contrast, the sorting scheme used by Drew et al. (2014), based on 236-taxon OV scores, was not effective in identifying such sites (as presented in Fig. 1 which shows the distance between angiosperms and gymnosperms estimated for equivalent B partition sites identified using the different OV scores and its ratio to the total tree length). Consequently, site removal based on the OV sorting protocol used by Drew et al. (2014) was unable to reduce LBA between out-group and in-group sequences.

When the 36-taxon OV scores were used to order the 236-taxon data matrix of Drew et al. (2014), the same pattern of eroding support for the basal placement of Amborella was observed (Fig. 2a) as occurred for less-densely sampled data sets (Goremykin et al. 2009, 2013). Similarity in changes to levels of support for alternative basal-most angiosperm branches with the 36-taxon and 236-taxon alignments (Fig. 2a) indicates that the increased sampling of crown group angiosperms in the latter does not improve reliability of angiosperm root placement.

Under the CAT+GTR+G model, removal of the 2250 positions that had the highest 36-taxon OV scores from alignment 1 resulted in a tree containing the basal-most ANT clade (Fig. 2b). This relationship was recovered with high (≥ 0.95) PP support for the next eight A partitions sampled (A217–A234). The ANT clade continued to be favored until 4000 of the most divergent positions were removed. As more sites were removed from the A partition (starting from A217), a strongly supported basal-most ANTI clade (Amborella plus Nymphaeales plus Illicium) became favored. When continuous gamma distribution, rather than a discrete four-category model, was used to accommodate positional rate heterogeneity, recovery of ANT and ANTI branches occurred at an earlier noise removal step. That is, a strongly supported basal-most ANT clade occurred for the A6–A16 partitions, and appearance of a strongly supported ANTI clade occurred for the A17–A34 partitions (Fig. 2c).

Reconstruction Accuracy with B Partition Data

We compared phylogenetic reconstruction accuracy for the B6 (1500 bp) partition and randomly sampled jackknife replicates of the same length. To quantify an error in different root placements obtained in the analyses of 1500 pos. long B6 partitions and jackknife replicates (and without favoring the Amborella-based hypothesis, the Nymphaeales-based hypothesis or the ANT hypothesis, which are currently discussed in the literature), we scored root placements as “potentially correct” when the outgroup branch was recovered at the sister group position to Amborella, or Nymphaeales, or Amborella plus Nymphaeales (areas shown in shades of green in Fig. 4), and alternative root placements as “erroneous” (shown in black and shades of gray in Fig. 4).

With the B6 partition data sampled from alignment 1 (ordered using the 36-taxon OV scores), 5 out of 50 Phylobayes trees were recorded as having a “potentially correct” rooting (two times resolved at Amborella branch and three times at the ANT branch). In the majority of trees (34), the gymnosperm branch was attached either to a large polytomy comprising all major angiosperm lineages or to branches subtending the mesangiosperms (Fig. 4a). With the jackknife resampled alignment 1 data, we recorded 35 “potentially correct” root placements from 50 analyses (Fig. 4b).

The above analyses were repeated for the 236-taxon data set sorted using the 36-taxon OV scores.
In 50 distinct Bayesian analyses of B6 partition, we recovered 4 “potentially correct” rooting, with out-group placements on the Amborella branch (Fig. 4c). In 38 trees, the gymnosperm outgroup was attached to branches subtending either monocots or eudicots (20 different attachments points). In four cases, the backbone of the angiosperm subtree was unresolved. In contrast, with 50 jackknife replicates from the S1 alignment (Drew et al. 2014), we recovered “potentially correct” angiosperm root placements 38 times (Fig. 4d).

We also observed that B6 partitions gave unexpected phylogenetic reconstructions under the GTR + G4 model. An LBA artifact, evidenced by appearance of Trithuria at the basal-most position among the angiosperms, was recovered in the optimal ML tree based on B6 partition sampled from alignment 1. In one single run by RAxML, this optimal tree was selected out of 100 ML trees built during the run. Another LBA artifact was the appearance of Centrolepis, a monocot in the order Poales at this position. This artifact was registered in the optimal ML tree selected by RAxML out of 100 ML trees built in one run from B6 partition sampled from the S1 alignment (Drew et al. 2014) sorted using the 36-taxon scores.

Our findings indicate that OV sorting with the 36-taxon scores: (i) was effective in identifying site patterns with evolutionary properties distinctly different from those that dominate the full alignment and (ii) raises concerns for the potential of B partition sites to mislead inference of angiosperm root placement.

**Parametric Simulations**

Simulations were conducted to examine phylogenetic reconstruction accuracy for OV-sorted alignment 1, its B16 partition and its A16 partition. Simulations for B16 partition under a realistic site-heterogeneous substitution model yielded parametric data set replicates for which phylogenetic recovery of the model trees under the tree selection criterion and model adopted by Drew et al. (2014) (RAxML plus GTR + G4 model) was low (36% for the Amborella basal-most model branch and 2% or less for other model branches) (Fig. 5a, details provided in Supplementary Table 2 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f). In contrast, recovery of the basal-most angiosperm relations in the model trees with data simulated under a CAT + GTR + G4 model for the conserved A16 partition of alignment 1 was close to 100% for all four substitution models tested (Fig. 5c), High recovery rates for all model trees in A16 partitions of 200 OV-sorted replicates (Supplementary Table 3 available on Dryad at http://dx.doi.org/10.5061/dryad.vg97f) indicate that recovery of the ANT branch is not an artifact related to an LBA induced by OV sorting in the absence of model mis-specification. The negative impact of adding B partition sites to the A16 partition was evidenced by a steep decrease (from 100% to 16% and less) in recovery rates for the basal most model angiosperm branches which consist of more than one OTU (Fig. 5b) in trees built from parametric replicates which simulated the full-alignment length.

**Site Saturation of Basal Angiosperms**

Given the disproportionately large distance between in-group and out-groups in the B partition, there will be a tendency for long-branched basal angiosperms to be drawn toward outgroup sequences in phylogenetic analyses of these sites. This problem is most significant for basal angiosperm sequences exhibiting the greatest site saturation with respect to other in-group taxa. A comparison was thus made of the relative extent of site saturation among basal angiosperm sequences (Amborella, Trithuria, and Nymphaea) on the Amborella most basal model tree. In the shortest partition studied (i.e., B1: the 250 most varied position interval), the length...
of the Amborella most basal model tree was 591 steps. Randomizing the Amborella sequence in this partition and then resoring the length of Amborella most basal tree on these data (100 randomized data sets) produced trees with a distribution of tree lengths ranging from 589 to 609 [598±4] steps. Repeating this analysis, but randomizing Trithuria instead of Amborella produced trees with a distribution of tree lengths from 598 to 620 [609±4] steps. Randomizing Nymphaea produced tree lengths ranging from 622 to 641 [633±4] steps. A similar trend was observed for all partitions lengths examined (250, 500, 1000, 1500, 2000, 2500 sites) between 250 and 2500 sites. That is, at the most varied sites identified by OV sorting, while none of the basal angiosperm sequences were convincingly completely random in any of the partitions examined, Amborella exhibited the highest level of site saturation in all of the partitions examined. Further, in unconstrained tree reconstructions, the randomized sequence always tended to be placed as sister to the other angiosperm sequences. This result, while not surprising, indicates the topological bias that is favored where there is site saturation of basal angiosperm sequences coupled with model mis-specification in phylogenetic reconstruction.

DISCUSSION

Data–Model Fit

Recent studies that have addressed the issue of the fit between model and data (Goremykin et al. 2013; Xi et al. 2014) find that Amborella is not the sole representative of an evolutionary lineage sister to other angiosperms. The central argument of Goremykin et al. (2013) was that systematic error could explain phylogenetic reconstructions that placed Amborella as distinct from all extant angiosperms. Here we report further observations on chloroplast sequence data which strengthen and elaborate this conclusion.

A first point of clarification is that better fitting site-heterogeneous models designed to counter LBA-related errors shift the low levels of support reported by Drew et al. (2014) for an Amborella most basal branch, to strong levels of support for the ANT relationship in analyses of first and second codon position data (sampled from the in-frame alignment presented in Goremykin et al. 2013). Second, reanalyses of concatenated nuclear, mitochondrial, and chloroplast sequence data from Solits et al. (2011) uniformly indicate that, contrary to the conclusion reported by the authors, these data support the ANT branch. A similar observation has also been made by Xi et al. (2014), who report that only the fastest evolving (saturated) sites in these data support an Amborella most basal hypothesis.

We show here that the fastest evolving sites in concatenated chloroplast sequences do not contribute to resolving basal angiosperm relationships in an informative way. Phylogenetic reconstruction for chloroplast B6 partition data (a subset of the fastest evolving sites) is much more error-prone than phylogenetic reconstruction for the same length randomly resampled full data (Fig. 4). Parametric simulations (Fig. 5) also demonstrate low reconstruction accuracy for B partition data. Xi et al. (2014) also report similar results for nuclear sequences. Removal of these sites from the chloroplast data matrix leads to a strongly supported ANT branch. As expected, less error-prone site heterogeneous models start to support the ANT branch when less number of saturated (Fig. 1a) and heterotachous (Fig. 1b) sites are discarded from the data matrix (Fig. 2). The problem, unrecognized by Drew et al. (2014), who advocate using total evidence approach to resolve basal angiosperm relationships, is that their data contain sites that contribute disproportionately to a very large distance between in-group and out-group. Such extreme heterotachy is not observed at the more conserved sites (Fig. 1). The very great evolutionary distance between in-group and out-group at these sites, coupled with a high degree of site saturation, draws Amborella toward the root and makes it appear as if it is sister to other extant angiosperms.

Taxon Sampling

Drew et al. (2014) suggested that the extent of their taxon sampling should give readers confidence in their conclusions concerning the most basal position of Amborella. However, no evidence has been provided for this speculation, and the argument does not receive support from observations reported elsewhere that suggest reduction in taxon sampling can also improve phylogenetic inference (e.g., Rokas et al. 2003; Gatesy et al. 2007). Angiosperm root inference is an example of a much discussed problem concerning attachment of a distantly related outgroup to a radiation (Whitfield and Lockhart 2007). Theoretical (Goldman 1998; Geuten et al. 2007) and simulation studies (Poe 2003; Townsend et al. 2010) emphasize the importance of taxon sampling that impacts most on the reconstruction accuracy of the deepest internal nodes. Dense sampling of highly derived taxa from within the ingroup (e.g., sampling crown group angiosperms as in Drew et al. 2014) can be expected to contribute little to resolution of deep internal nodes (Graybeal 1998; Poe 2003; Mossel and Steel 2005; Townsend et al. 2010). This is also the case with angiosperm root inference. Our analyses of conserved (Fig. 2a) and variable sites (Fig. 4) do not support the view that very dense taxon sampling of eudicots and monocots increases phylogenetic accuracy of angiosperm root placement.

Sorting of Site Patterns

There are both tree-dependent and tree-independent approaches to sort site patterns in terms of substitution rate. While the former group of methods is not free from systematic error in estimation of site-specific substitution rates due to a wrong tree, the latter group is. One of
the most computationally simplest, but most effective tree-independent methods is OV sorting (Goremykin et al. 2010). However, this and other sorting protocols need to be applied cautiously, particularly when the taxon number is large (see discussions in Goremykin et al. 2010, p. 324; Mosel and Roch 2013). This issue was recently highlighted by the findings of Drew et al. (2014) who, using a different taxon sampling scheme, obtained results that differed from those obtained by Goremykin et al. (2013). They found that appearance of the ANT branch occurred only after many more sites were stripped from a concatenated chloroplast data set.

Their result can be easily explained. In Drew et al.’s (2014) S1 alignment, mono- and eudicots together constituted more than 95% of the angiosperm taxon set. Consequently, their OV scores (based on 236 OTUs) reflected overall character variability, mostly among crown group angiosperms and did not concentrate sites that were saturated between in-group and out-group toward the end of the concatenated alignment (Fig. 1), which made identification of LBA artifacts unlikely. In contrast, when OV scores were calculated based on a 36-taxon set—that is, one that does not include 200 crown group angiosperms—sites that were saturated between angiosperms and gymnosperms were easily identified (Fig. 1). Drew et al.’s results are not indicative of an intrinsic failure of the OV method, but rather of the taxon sampling scheme used which masked the extent and impact of saturation between in- and out-group (Fig. 1a), and which was applied contrary to recommendations made for application of the method (Goremykin et al. 2010).

When a small proportion of the fastest evolving characters which have high levels of site saturation between gymnosperms and angiosperms are removed, the data sets of Goremykin et al. (2009, 2013) and Drew et al. (2014) all yield ANT as a basal clade. Exclusion of Gnetales and nonvascular plants, contrary to the speculation of Drew et al. (2014), does not affect this result (Fig. 2). Support for the ANT clade after removal of a small proportion of such sites is, thus, a general phenomenon, robust to: (i) changes of taxon sampling in in- and out-groups, (ii) changes in gene sampling, and (iii) different alignment construction strategies and editing applied in the above studies.

Do We have a Definitive Answer for Basal Angiosperm Relationships?

Our parametric simulations (Fig. 5) suggest that the conserved A partition for concatenated chloroplast sequences provides reliable information for inferring relationships among basal angiosperms. This was not the case for B16 partition data. Nor was it the case if B relationships among basal angiosperms. This was not the case for the ANT clade after removal of a small proportion of sites that were saturated between in-group and out-group toward the end of the concatenated alignment (Fig. 1), which made identification of LBA artifacts unlikely. In contrast, when OV scores were calculated based on a 36-taxon set—that is, one that does not include 200 crown group angiosperms—sites that were saturated between angiosperms and gymnosperms were easily identified (Fig. 1). Drew et al.’s results are not indicative of an intrinsic failure of the OV method, but rather of the taxon sampling scheme used which masked the extent and impact of saturation between in- and out-group (Fig. 1a), and which was applied contrary to recommendations made for application of the method (Goremykin et al. 2010).

When a small proportion of the fastest evolving characters which have high levels of site saturation between gymnosperms and angiosperms are removed, the data sets of Goremykin et al. (2009, 2013) and Drew et al. (2014) all yield ANT as a basal clade. Exclusion of Gnetales and nonvascular plants, contrary to the speculation of Drew et al. (2014), does not affect this result (Fig. 2). Support for the ANT clade after removal of a small proportion of such sites is, thus, a general phenomenon, robust to: (i) changes of taxon sampling in in- and out-groups, (ii) changes in gene sampling, and (iii) different alignment construction strategies and editing applied in the above studies.

Do We have a Definitive Answer for Basal Angiosperm Relationships?

Our parametric simulations (Fig. 5) suggest that the conserved A partition for concatenated chloroplast sequences provides reliable information for inferring relationships among basal angiosperms. This was not the case for the ANT clade after removal of a small proportion of sites that were saturated between in-group and out-group (Fig. 1a), and which was applied contrary to recommendations made for application of the method (Goremykin et al. 2010).

When a small proportion of the fastest evolving characters which have high levels of site saturation between gymnosperms and angiosperms are removed, the data sets of Goremykin et al. (2009, 2013) and Drew et al. (2014) all yield ANT as a basal clade. Exclusion of Gnetales and nonvascular plants, contrary to the speculation of Drew et al. (2014), does not affect this result (Fig. 2). Support for the ANT clade after removal of a small proportion of such sites is, thus, a general phenomenon, robust to: (i) changes of taxon sampling in in- and out-groups, (ii) changes in gene sampling, and (iii) different alignment construction strategies and editing applied in the above studies.

Do We have a Definitive Answer for Basal Angiosperm Relationships?

Our parametric simulations (Fig. 5) suggest that the conserved A partition for concatenated chloroplast sequences provides reliable information for inferring relationships among basal angiosperms. This was not the case for the ANT clade after removal of a small proportion of sites that were saturated between in-group and out-group (Fig. 1a), and which was applied contrary to recommendations made for application of the method (Goremykin et al. 2010).

When a small proportion of the fastest evolving characters which have high levels of site saturation between gymnosperms and angiosperms are removed, the data sets of Goremykin et al. (2009, 2013) and Drew et al. (2014) all yield ANT as a basal clade. Exclusion of Gnetales and nonvascular plants, contrary to the speculation of Drew et al. (2014), does not affect this result (Fig. 2). Support for the ANT clade after removal of a small proportion of such sites is, thus, a general phenomenon, robust to: (i) changes of taxon sampling in in- and out-groups, (ii) changes in gene sampling, and (iii) different alignment construction strategies and editing applied in the above studies.


