Systematic Biology Advance Access published May 20, 2015

The root of flowering plants and total evidence

Authors: GOREMYKIN[1*], V.V., NIKIFOROVA[1], S.V., CAVALIERI[1], D., PINDO[1], M. AND PETER LOCKHART[2,3]

[1]*FEM Research and Innovation Center, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy;*

[2]*Institute of Molecular Biosciences, Massey University, Palmerston North, New Zealand;*

[3]*Institute 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

phone: +39 0461 615 647

Keywords. Systematic error, model fit, parametric simulations, total evidence approach, angiosperm origins, *Amborella*

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 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 artefact (LBA) 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 finding that the basal angiosperm *Amborella* is most closely related to aquatic herbs. Our current and previously reported (Goremykin et al. 2009, 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.

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; Soltis 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 analysed (Barkman et al. 2000; Soltis et al. 2000; 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; Mardanov et al. 2008; Moore et al. 2007; Wu et al. 2007; 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 artefact due to systematic error and a poor fit between time reversible substitution models and sequence data. The same paper 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 reconstructions 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 that 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 OV 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 misspecification 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 misled 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; Soltis 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. That 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 were 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; Kolaczkowski and Thornton, 2004; Schwarz et al.2004; Thornton and Kolaczkowski, 2005; Spencer et al. 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; Burleigh and Mathews 2007; Pisani 2004; Pisani et al. 2012; Delsuc et al. 2005; Philippe et al. 2005; Jeffroy et al. 2006; Rodriguez-Ezpeleta et al. 2007; Sperling et al. 2009; Cummins and McInerney 2011, 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; Qiu 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 ingroup and outgroup 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 datasets studied by Goremykin et al. (2013) are similar to the evolutionary properties of the datasets 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. This matrix uploaded by Drew et al. (2014) to the Dryad site is 58950 pos. long.

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 dataset 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 58554 pos. long alignment of 36 OTUs is henceforth referred to as alignment 1 (suppl. file 1).

Alignment 2: 1&2 codon position Data Set from Goremykin et al. (2013)

In order 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 (21116 pos. long alignment 2, suppl. file 2).

Alignment 3: Reduced 57 OTU Data Set from Soltis et al. (2011)

We reexamined the Soltis et al. (2011) data. To conduct time-consuming Bayesian analyses with more appropriate substitution models, we reduced this 640-taxon matrix to an alignment of 57 OTUs which contained all 8 members of the gymnosperm outgroups and 49 angiosperm species, including all basal angiosperms present in the Soltis et al. (2011) data set. The reduced alignment of 57 OTUs is referred to as alignment 3 (suppl. file 3).

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 covarion model implemented in Phylobayes) models**,** and on alignment 3, specifying GTR+G, CAT+GTR+G, and CAT+GTR+G+covext models. Five chains were run under each model/data combination until 2000 cycles were sampled. For

every chain, we have discarded the first 500 cycles as “burn-in” which was sufficient for all chains to reach maximized likelihood values, and sampled every cycle thereafter to build trees.

OV Scores Used to Sort Datasets 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 suppl. file 4) was divided into bi-partitions: 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 x 250 (where n=1, 2, 3...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. 2010, 2013). Each A_n partition contained 58554-250 x *n* sites and each B_n partition contained 250 x *n* sites.

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 ingroup and the outgroup.

For this analysis, we took the unedited 236-taxon S1 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 S1 alignment sorted with 36-taxon OV scores (provided as suppl. file 5) and from the S1 alignment sorted with 236-taxon OV scores (provided as suppl. file 6). Trees were built for the

first 24 B partitions of the resulting two sorted alignments using RaxML and the GTR+G_4 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 gymnosperm outgroup and ii) the proportion of the branch subtending the outgroup 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 (1) and 236-taxon (S1) 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).

A site-homogeneous model and maximum likelihood heuristic was used in the above comparisons because reaching convergence under more realistic CAT-based models required impractically long run times for the genome-scale 236-taxon S1 alignment and its larger B partitions. However, we did undertake phylogenetic analyses of alignment 1, assuming a CAT substitution model and gamma distribution of site-specific rates approximated by four discrete categories. For this analysis, we ran unconstrained chains under the CAT+GTR+G_4 model for sorted alignment 1 and its $A_{1\text{-}24}$ partitions, sampling for 2000 cycles. We discarded 500 cycles as burn-in and built trees, sampling every cycle thereafter and registering changes in posterior probability (PP) support for alternative basal-most angiosperm clades (Fig. 2b). We

checked for the effect of approximating the gamma distribution with four rate classes by repeating the above analyses (Fig. 2c), specifying continuous gamma distribution (CAT+GTR+G model) and comparing bootstrap support for basal angiosperm relationships.

Four different unweighted trees were identified with the above analyses (Fig. 3, Suppl. Fig. 1). 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 1. Since trees built from the conserved A partition under a CAT+GTR+G model begin to favor *Amborella* grouping with *Trithuria* and Nymphaeales (i.e., the ANT clade) with maximum PP support (Fig. 2c) from the sixth

shortening (B_6) step on (i.e., after removal of at least 1500 sites with the highest OV scores), we built trees for the residual B_6 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 B_6 partition of alignment 1, each time sampling 20,000 cycles and discarding the first 10,000 cycles to build trees. The very large cut-off of 10,000 cycles, reaching far into a plateau of maximized likelihood scores, was chosen to highlight the absence of convergence of mature chains. Different attachments of the gymnosperm outgroup registered in these experiments are shown in Fig. 4 (Y-axis, experiment A). In parallel, we also analyzed 50 non-overlapping 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 B_6 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 B_6 partition, and built trees after discarding the first 35,000 cycles. Again, this cut-off, reaching far into a plateau of maximized likelihood scores, was chosen to highlight the absence of convergence

of mature chains. The results of these experiments are presented in Fig. 4 as experiment C. We also analyzed 50 jackknife replicates of the S1 alignment which were of the same length as the B_6 partition (1500 pos.). These replicates were obtained using the Seqboot program. 20,000 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 burn-in (Fig. 4, Y-axis, experiment D). A more detailed description of experiments summarized in Fig. 4 is presented in supplementary table 1.

We also conducted RAxML analyses, and used the same substitution model adopted by Drew et al. (GTR+G_4) in order to identify the best-scoring ML tree among 100 ML trees built starting from 100 randomized MP trees for the B_6 partitions analyzed above (36-taxon and 236-taxon data matrices).

Parametric Simulations

We used parametric simulations to evaluate the accuracy of the tree reconstruction method from full-length data (alignment 1) and its A and B partitions. For reconstructions, we used RAxML and assumed a GTR+G_4 model as used by Drew et al. (2014). Simulations assumed a CAT+GTR+G_4 model which was previously found in cross-validation experiments (Goremykin et al. 2013) to provide a good fit for a dataset of concatenated chloroplast protein-coding genes. We simulated 36-taxon sequence alignments with Phylobayes based on alignment 1 and its A_{16} and B_{16} partitions that had been created using the sorter.pl script. The sixteenth shortening step was chosen because, beginning with this shortening step, a basal-most *Amborella* plus *Trithuria* plus Nymphaeales clade was consistently

recovered with high (BP >=80%) support under the GTR+G_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+GTR+G_4 model for the A_{16} and B_{16} partitions and the full data set. For these simulations, four model tree topologies (Fig. 3, Suppl. Fig. 1) 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 (-s 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 model parameters.

In phylogenetic analyses with a GTR+G_4 model (the reconstruction model used by Drew et al. 2014), we evaluated whether or not we could recover the correct placement of outgroups in the model trees. We reconstructed trees from 600 data matrix replicates, simulated for the full length and partitioned observed data. We recorded the percentage of times RaxML recovered correct and spurious attachments of the gymnosperm branch to the angiosperm subtree in these experiments (Fig. 5). Detailed outcomes for these analyses are summarized in supplementary Table 2.

In order 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 gymnosperm outgroup to the angiosperm subtree from the 54554 pos. long A_{16} partitions of the sorted replicates (shown in Suppl. Table 3).

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 are 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– 2,500 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 2x2 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.

Win-Paup4b10 (Swofford 2002) was used to calculate the parsimony scores and also the topologies of unconstrained parsimony trees.

RESULTS

Soltis et al. (2011) data and the 1st and 2nd codon positions from in-frame alignment of Goremykin et al. (2013)

All 10 Bayesian analyses of 1st and 2nd codon positions of our previously published in-frame alignment (Goremykin et al. 2013) using the CAT+GTR+G and CAT+GTR+G+covext models yielded trees that contained the basal-most ANT clade (Suppl. Fig. 2a). The clade was well supported both under the CAT+GTR+G+covext 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 ten experiments (five per model) (Suppl. Fig. 2b).

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 outgroup, 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 Figure 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 outgroup and ingroup 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_4 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 (A_{10} – A_{16}). 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 A_{18}), a strongly supported basal-most ANTI clade (*Amborella* plus Nymphaelaes 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 A_6 – A_{16} partitions, and appearance of a strongly supported ANTI clade occurred for the A_{17} – A_{24} partitions (Fig. 2c).

Reconstruction accuracy with B partition data

We compared phylogenetic reconstruction accuracy for the B_6 (1500 bp) partition and randomly sampled jackknife replicas of the same length. In order to quantify an error in different root placements obtained in the analyses of 1500 pos. long B_6 partitions and jackknife replicates (and without favoring the *Amborella*-basal hypothesis, the Nymphaeales-basal 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 B_6 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 B_6 partition, we recovered four “potentially correct” rootings, with outgroup 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 B_6 partitions gave unexpected phylogenetic reconstructions under the GTR+G_4 model. An LBA artefact, evidenced by appearance of *Trithuria* at the basal-most position among the angiosperms, was recovered in the optimal ML tree based on B_6 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 artefact was the appearance of *Centrolepis*, a monocot in the order Poales at this position. This artefact was registered in the optimal ML tree selected by RaxML out of 100 ML trees built in one run from B_6 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 alignment 1, its B_{16} partition and its A_{16} partition. Simulations for B_{16} partition under a realistic site-heterogeneous substitution model yielded parametric dataset 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+G_4 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). In contrast, recovery of the basal-most angiosperm relations in the model trees with data simulated under a CAT+GTR+G_4 model for the conserved A_{16} partition of alignment 1 was close to 100% for all four substitution models tested (Fig. 5c). High recovery rates for all model trees in A_{16} partitions of 200 OV-sorted replicates (Suppl. Table 3) indicate that recovery of the ANT branch is not an artefact 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 A_{16} 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 ingroup and outgroups in the B partition, there will be a tendency for long branched basal angiosperms to be

drawn towards 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 ingroup 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 rescoring the length of *Amborella* most basal tree on these data (100 randomised data sets) produced trees with a distribution of tree lengths ranging from 589-609 [598+/-4] steps. Repeating this analysis, but randomizing *Trithuria* instead of *Amborella* produced trees with a distribution of tree lengths from 598-620 [609+/-4] steps. Randomizing *Nymphaea* produced tree lengths ranging from 622-641 [633+/-4] steps. A similar trend was observed for all partitions lengths examined (250, 500, 1000, 1500, 2000, 2500 sites) between 250-2500 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, whilst not surprising, indicates the topological bias that is favoured 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). Secondly, re-analyses of concatenated nuclear, mitochondrial and chloroplast sequence data from Soltis 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 B_6 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 (Fig1a) and heterotachous (Fig2b) 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 ingroup and outgroup. Such extreme heterotachy is not observed at the more conserved sites (Fig. 1). The very great evolutionary distance between ingroup and outgroup at these sites, coupled with a high degree of site saturation, draws *Amborella* towards 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; Townsend 2007) and simulation studies (Poe 2003; Townsend and López-Giráldez 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 and López-Giráldez 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 and also Mossel 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 dataset.

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 ingroup and outgroup towards the end of the concatenated alignment (Fig. 1), which made identification of LBA artefacts unlikely. In contrast, when OV scores were calculated based on a 36-taxon set— i.e. 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 outgroup (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 datasets of Goremykin et al. (2009, 2013) and Drew et al. (2014) all yield ANT as a basal clade. Exclusion of Gnetales and non-vascular 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 outgroups, 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 B_{16} partition data. Nor was it the case if B partition sites were added to conserved A partition sites. The reconstruction accuracy for correct root placement substantially decreased when B partition sites were added under three of the four evolutionary models used. This finding indicates that model-violating sites have the potential to mislead not only analyses of B partition data, but also analyses of full-length sequence data. This places a different interpretation on the recent findings of Drew et al. (2014). They reported that same basal placement for *Amborella* for B partition and A & B partition data combined, and concluded this root placement must, therefore, be correct. However, our findings suggest that their phylogenetic inferences for both B and A partitions are likely to be incorrect. The presence of model mis-specified sites in both data sets misleads their inference of the basal-most model relationships.

Had the *Amborella* basal-most hypothesis been truly supported by the data, we would expect its recovery with *both* alignment 1 and its A_{16} partition. This was not observed. Support for the *Amborella* most basal hypothesis in observed data comes from B partition sites that have poor qualities for tree building. Most importantly, our randomization analysis indicates that the B partition sites with greatest character state variation are likely to contribute to LBA between *Amborella* and the outgroup.

Had the ANTI and ANT branches been supported by the data, we would expect their recovery from the A_{16} partition, but not from the full data set. Because we registered no support for the ANTI branch from the A_{16} partition of alignment 1 (Fig. 2), we currently consider the ANT hypothesis to be the best hypothesis for

basal angiosperm relationships based on our analyses. This hypothesis is also the one most supported in analyses of nuclear data (Xi et al. 2014).

Concluding, it should be noted that the predictive power of simulation tests presented here depends on how well the model chosen to generate the replicate data describes the substitution process in the observed data. In the case of the perfect (unattainable) fit, the accuracy of the phylogeny reconstruction method can be evaluated more precisely. In the future, with development of non-time reversible models (e.g. Jayaswal et al, 2014) the predictive power of such tests should only increase, at which point it will be interesting to update the reliability of the basal-most angiosperm relationships. It should be noted that that taxon sampling among basal angiosperms is sparse in all data sets, and the significance of this should not be underestimated for phylogenetic reconstructions that seek to assign ancestral traits to the earliest angiosperms.

SUPPLEMENTARY MATERIAL AVAILABLE ON THE DRYAD SITE

SA1 archive file containing all supplementary materials mentioned in the text.

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.

Ané, C., Burleigh, J. G., McMahon, M. M., Sanderson, M. J. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol. Biol. Evol. 22:914–24.

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. 97:13166–71.

Bausher, M. G., Singh, N. D., Lee, S.-B., Jansen, R. K., Daniell, H. 2006. The complete chloroplast genome sequence of Citrus sinensis (L.) Osbeck var "Ridge Pineapple": organization and phylogenetic relationships to other angiosperms. BMC Plant Biol. 6:21.

Brinkmann, H., Philippe, H. 1999. Archaea sister group of bacteria? Indications from tree reconstruction artefacts in ancient phylogenies. Mol. Biol. Evol. 16:817–25.

Burleigh, J. G., Mathews, S. 2007. Assessing systematic error in the inference of seed plant phylogeny. Internat. J. Plant Sci.168:125–135.

Carnap, R. 1950. Logical Foundations of Probability. Chicago: University of Chicago Press.

Chang, J.T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. Math. Biosci. 134: 189–215.

Chang, C., Lin, H., Lin, I., Chow, T., Chen, H-H., Chen, W-H., Cheng, C-H., Lin, C-Y., Liu, S-M., Chang, C-C., Chaw, S. 2005. The Chloroplast Genome of *Phalaenopsis aphrodite* (Orchidaceae): comparative analysis of evolutionary

rate with that of grasses and Its phylogenetic implications. Mol. Biol. Evol. 23:279–291.

Chase, M.W., Duvall, M.R. Hills, H.G. Conran, J.G. Cox, A.V. Eguiarte, L.E., Hartwell, J., Fay, M.F., Caddick, L.R., Cameron, K.M., Hoot. S. 1995. Molecular phylogenetics of Lilianae. In P. J.Rudall, P. J. Cribb, D. F. Cutler, and C. J. Humphries [eds.], Monocotyledons: systematics and evolution, 109–137. Royal Botanic Gardens, Kew, UK.

Chaw, S. M., Parkinson, C. L., Cheng, Y., Vincent, T. M., Palmer, J. D. 2000. Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proc. Natl. Acad. Sci.U.S.A. 97:4086–91.

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. 58:424–41.

Cooper, E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends in Plant Science 19(9): 576-582.

Cummins, C., McInerney. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60:833.

Delsuc, F., Brinkmann, H., Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Rev. Genet. 6:361–75.

Drew, B. T., Ruhfel, B. R., Smith, S. A., Michael, J., Briggs, B. G., Gitzendanner, M. A., Soltis, P.E., Soltis, D. E. 2014. Another look at the root of the angiosperms reveals a familiar tale. Syst. Biol. 63:368–382.

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

Gatesy, J., DeSalle, R., Wahlberg, N. 2007. How many genes should a systematist sample? conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Syst. Biol. 56:355–363.

Geuten, K., Massingham, T., Darius, P., Smets, E., Goldman, N. 2007. Experimental design criteria in phylogenetics: where to add taxa. Syst. Biol. 56:609–622.

Goldman, N. 1998. Phylogenetic information and experimental design in molecular systematics. Proc. Royal Society B-Biological Sciences. 265:1779–1786.

Goremykin, V., Bobrova, V., Pahnke, J., Troitsky, A., Antonov, A., Martin, W. 1996. Noncoding sequences from the slowly evolving chloroplast inverted repeat in addition to *rbc*L data do not support Gnetalean affinities of angiosperms. Mol. Biol. Evol. 13:383–96.

Goremykin, V. V, Hirsch-ernst, K. I., Wo, S., Hellwig, F. H. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. Mol. Biol. Evol. 363:1499–1505.

Goremykin, V. V, Viola, R., Hellwig, F. H. 2009. Removal of noisy characters from chloroplast genome-scale data suggests revision of phylogenetic placements of *Amborella* and *Ceratophyllum*. J. Mol. Evol. 68:197–204.

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.

Goremykin, V. V, Nikiforova, S. V, Biggs, P. J., Zhong, B., Delange, P., Martin, W., Woetzel, S., Atherton, R.A., McLenachan, P.A., Lockhart, P. J. 2013. The evolutionary root of flowering plants. Syst. Biol. 62:50–61.

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

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

Hansmann, S., Martin, W. 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. Internatl. J. Syst. Evol. Micro. 50:1655–1663.

Hasegawa, M., Hashimoto, T. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hilu, K.W., Borsch, T., Muller, K., Soltis, D.E., Soltis, P.S., Savolainen, V., Chase, M.W., Powell, M.P., Alice, L.A, Evans, R., Sauquet, H., Neinhuis, C., Slotta,

T.A.B., Jens, G.R., Campbell, C.S., Chatrou, L.W. 2003 Angiosperm phylogeny based on matK sequence information. Am. J. Bot. 90:1758–1776.

Hirt, R.P., Logsdon, J.M. Jr., 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. USA 96:580–585.

Jansen, R.K., Kaittanis, C., Saski, C., Lee, S.-B., Tomkins, J., Alverson, A. J., Daniell, H. 2006. Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol. Biol. 6:32.

Jansen, R.K., Cai, Z., Raubeson, L.A., Daniell, H., Depamphilis, C.W., Leebens-Mack, J., Müller, 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. Nat. Acad. Sci. USA 104:19369-74.

Jayaswal, V., Wong, T.K., Robinson, J., Poladian, L., Jermiin, L.S. 2014. Mixture Models of Nucleotide Sequence Evolution that Account for Heterogeneity in the Substitution Process Across Sites and Across Lineages. *Syst. Biol.* 63(5):726–742.

Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of incongruence? Trends in Genet. 22(4):225–31.

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. E., 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.

Källersjö, M., Albert, V.A., Farris, J.S. 1999. Homoplasy increases phylogenetic structure Cladistics, 15: 91-93.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Systematic Zoology 38: 7–25.

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

Lanave, C., Preparata, G., Saccone, C., Serio, G. 1984 A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20: 86-93.

Lartillot, N., Philippe, H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. Phil. Trans. Royal Soc. B-Biol. Sci. 363:1463–1472.

Laurin-Lemay, S., Brinkmann, H., Philippe, H. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Current Biology 22:R593–4.

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. 22:1948-63.

Lockhart, P. J., Steel, M. A, Hendy, M. D., Penny, D. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11: 605–12.

Lockhart, P. J., Larkum, A. W., 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. USA 93:1930–4.

Lockhart, P., Steel, M. 2005. A tale of two processes. Syst. Biol. 54:948–951.

Lockhart, P., Novis, P., Milligan, B. G., Riden, J., Rambaut, A., Larkum, T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol. Biol. Evol. 23:40–5.

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.

Mardanov, A. V, Ravin, N. V, Kuznetsov, B. B., Samigullin, T. H., Antonov, A. S., Kolganova, T. V, Skyabin, K. G. 2008. Complete sequence of the duckweed (*Lemna minor*) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms. J. Mol. Evol. 66:555–64.

Moore, M. J., Bell, C. D., Soltis, P. S., Soltis, D. E. 2007. Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proc. Natl. Acad. Sci. USA 104:19363–8.

Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G., Soltis, D. E. 2010. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. Proc. Natl. Acad. Sci. USA 107:4623–4628 .

Mossel, E., Steel, M. 2005. How much can evolved characters tell us about the tree that generated them? In O.Gascuel (Ed.), In Mathematics of Evolution and Phylogeny (pp. 384–412). Oxford University Press.

Mossel, E., Roch, S. 2013. Identifiability and inference of non-parametric rates-across-sites models on large-scale phylogenies. J. Math. Biol. 67:767-797.

Penny, D., Hendy, M. D., Steel, M. A. 1992. Progress with Methods for Constructing Evolutionary Trees. Trends in Ecol. Evol. 7:73–79.

Philippe, H., Delsuc, F., Brinkmann, H., Lartillot, N. 2005. Phylogenomics. Ann. Rev. Ecol. Evol. Syst. 36:541–562.

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

Pisani, D., Feuda, R., Peterson, K. J., Smith, A. B. 2012. Resolving phylogenetic signal from noise when divergence is rapid: A new look at the old problem of echinoderm class relationships. Mol. Phyl. Evol. 62:27–34.

Poe, S. 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52:423–428.

Qiu, Y.-L., Dombrovska, O., Lee, J. H., Li, L. B., Whitlock, B. A., Bernasconi Quadroni, F., Rest, J. S., Davis C. C., Borsch, T., Hilu K. W., Renner, S. S.,

Soltis, D. E., Soltis, P. S., Zanis, M. J., Cannone, J. J., Gutell, R. R., Powell, M., Savolainen, V., Chatrou, L. W., Chase, M. W. 2005. Phylogenetic analyses of basal angiosperms based on nine plastid, mitochondrial, and nuclear genes. Int. J. Plant Sci. 166: 815 – 842.

Qiu, Y-L., Li, L., Hendry, T. A., Li, R., Taylor, D. W., Issa, M. J., Ronen, A. J., Vekaria, M. L., White, A. M. 2006. Reconstructing the basal angiosperm phylogeny : evaluating information content of mitochondrial genes. Taxon 55:837–856.

Qiu, Y.-L., Li, 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.

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.

Rokas, A., Williams, B. L., King, N., Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ruiz-Trillo, I., Riutort, M., Littlewood, D., Herniou, E., Baguna, J. 1999. Acoel Flatworms : Earliest Extant Bilaterian Metazoans, Not Members of Platyhelminthes. Science 283:1919–1923.

Savolainen, V., Chase, M.W, Hoot, S.B., Morton, C.M, Soltis, D.E, Bayer C. (2000) Phylogenetics of flowering plants based on combined analysis of plastid atpB and rbcL gene sequences Syst. Biol. 49: 306-362.

Schwarz, M.P., Tierney, S.M., Cooper, S.J.B., Bull, N.J. 2004. Molecular phylogenetics of the allodapine bee genus Braunsapis: A/T bias and heterogeneous substitution parameters. Mol. Phylogenet. Evol. 32: 110–122.

Soltis, P. S., Soltis, D. E., Zanis, M. J., Kim, S. 2000. Basal Lineages of Angiosperms : Relationships and implications for floral evolution. Internatl. J. Plant Sci. 161:S97–S107.

Soltis, D.E., Albert, V.A, Savolainen, V., Hilu, K., Qiu, Y.-L., Chase, M.W., Farris, J. S., Stefanovic, S., Rice, D.W., Palmer, J.D. Soltis, P.S. 2004. Genome-scale data, angiosperm relationships, and “ending incongruence”: a cautionary tale in phylogenetics. Trends in Plant Sci. 9:477–83.

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. Amer. J. Bot. 98:704–730.

Spencer, M., Susko, E., Roger, A.J. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22: 1161–1164.

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.

Steel, M. A., Lockhart, P. J., Penny, D. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

Steel, M., Lockhart, P. J., Penny, D. 1995. A frequency-dependent significnace test for parsimony. Mol. Phyl. Evol. 4:64–71.

Stefanović, 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.

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

Thornton, J.W., Kolaczkowski, B. 2005. No magic pill for phylogenetic error. Trends Genet. 21: 310–311.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Townsend, J., López-Giráldez, F. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59:446-457.

Waddell, P. J. 2005. Measuring the fit of sequence data to phylogenetic model: allowing for missing data. Mol. Biol. Evol. 22:395–401.

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

Wodniok, S., Brinkmann, H., Glöckner, G., Heidel, A. J., Philippe, H., Melkonian, M., Becker, B. 2011. Origin of land plants: do conjugating green algae hold the key? BMC Evol. Biol. 11:104.

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. 24:1366–79.

Xi, Z., Liu, L., Rest, J.S., Davis, C.C. 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. DOI:10.1093/sysbio/syu055

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. USA 99:6848–53.

FIGURE LEGENDS

FIGURE 1

A) Lengths of branches connecting angiosperm subtrees to outgroups 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 (red dotted line) and applying OV scores estimated based on the 36-taxon subset (blue dotted line). 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 % of the total length of trees estimated from B partitions obtained applying OV scores computed based on all 236 taxa (red dotted line) and obtained based on OV scores for the 36 taxa (blue dotted line). Values on the Y axis indicate % values and values on the X axis indicate length of corresponding B partitions.

FIGURE 2

A) Bootstrap support values obtained for alternative basal-most angiosperm branches in unrooted RAxML trees built under a GTR+G_4 model based on 36-taxon alignment 1 and it's 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 it's 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) Posterior probability support obtained for alternative basal-most angiosperm branches and alternative sister groups to angiosperms in unrooted Phylobayes trees

built under CAT+GTR+G_4 model based on 36-taxon alignment 1 and it's A partitions. Values on the Y axis indicate the posterior probability values and values on the X axis indicate length of the corresponding B partitions.

C) Posterior probability 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 it's A partitions. Values on the Y axis indicate the posterior probability values and values on the X axis indicate length of the corresponding B partitions.

FIGURE 3

Schematic representation of fully resolved, unrooted cladograms recovered in analyses of A partitions of 36-taxon alignment 1, that were used as model trees for simulation analyses. Fig. 3a, 3b, 3c and 3d show alternative hypotheses of phylogenetic relationships among basal angiosperms and gymnosperms consecutively recovered as more divergent sites were removed (Fig. 2).

FIGURE 4

Gymnosperm outgroup placements recovered in unrooted optimal trees obtained in 50 parallel Bayesian analyses for the 1500 pos. long B_6 partition of alignment 1 (shown on the Y-axis as experiment A), from non-overlapping 1500 pos. long jackknife replicates sampled from the full length alignment 1(shown on the Y-axis as experiment B), from the 1500 pos. long B_6 partition of 236-taxon S1 alignment based on 36-taxon OV scores (shown on the Y-axis as experiment C), and from non-overlapping 1500 pos. long jackknife replicates sampled from the full length S1 alignment (shown on the Y-axis as experiment D). All analyses assumed a

CAT+GTR+G model. Alternative basal-most angiosperm branches commonly recovered in recently published analyses are indicated by shades of green while unexpected placements are indicated by gray shading.

FIGURE 5

Recovery rates for basal-most angiosperm relationships in unrooted trees built from parametric replicates generated for alignment 1, its A_{16} and B_{16} partitions assuming CAT+GTR+G_4 substitution model under four evolutionary constraints – A (*Amborella* basal-most, Fig. 3a), ANTC (basal-most branch subtending *Amborella* plus Nymphaeales s.s. plus *Trithuria,* with Cycas sister to angiosperms, Fig. 3b), ANTCG (basal-most branch subtending *Amborella* plus Nymphaeales s.s. plus *Trithuria*, with the branch subtending Cycas+Ginkgo as sister to angiosperms, Fig. 3c) and ANTI (basal-most branch subtending *Amborella* plus Nymphaeales s.s. plus *Trithuria* plus *Illicium*, with the branch subtending Cycas+Ginkgo as sister to angiosperms, Fig. 3d). Rates of recovery of correct branches are shown in green, and rates of erroneous identification are shown in black and shades of gray.

Fig. 1

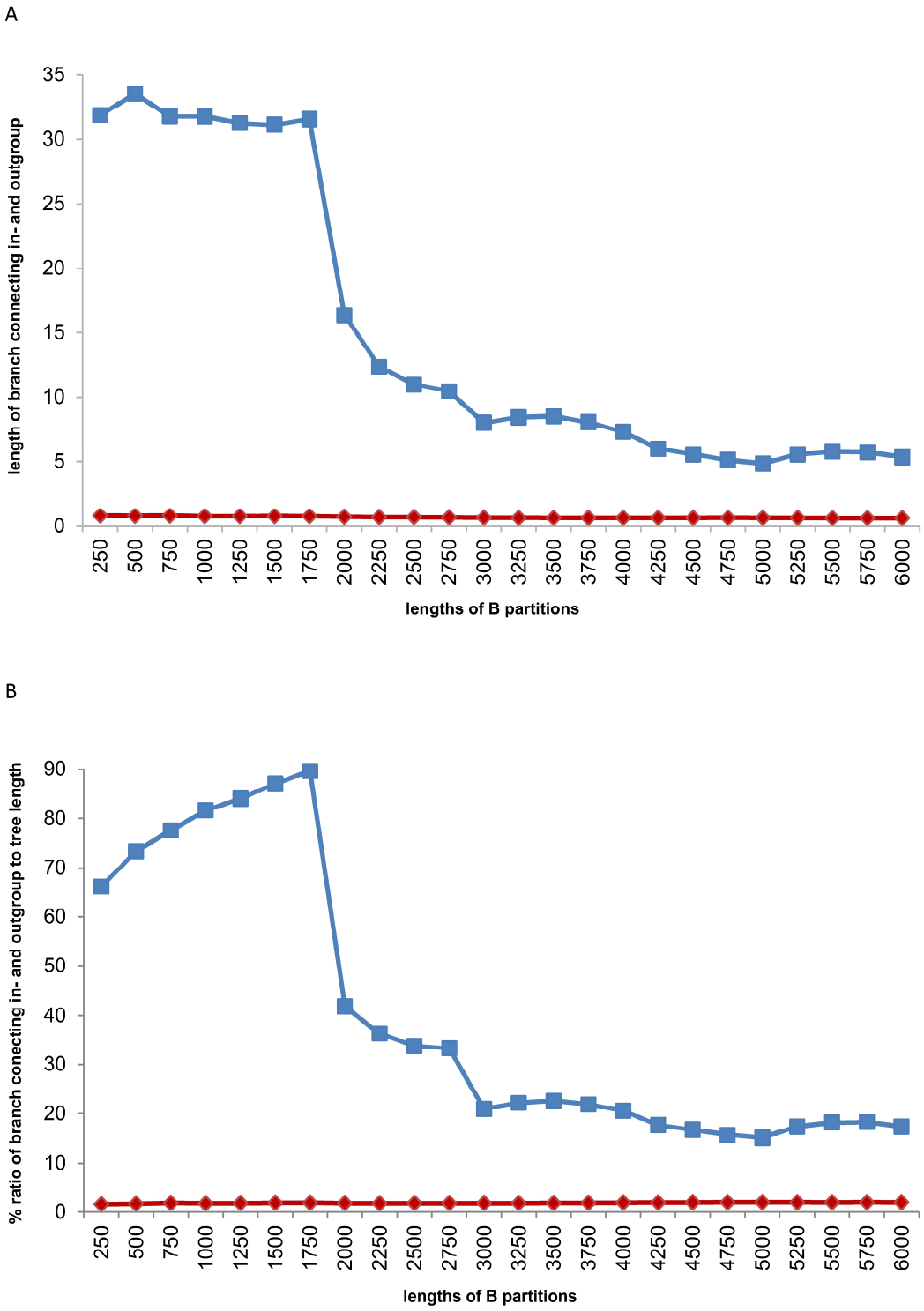

Fig. 2

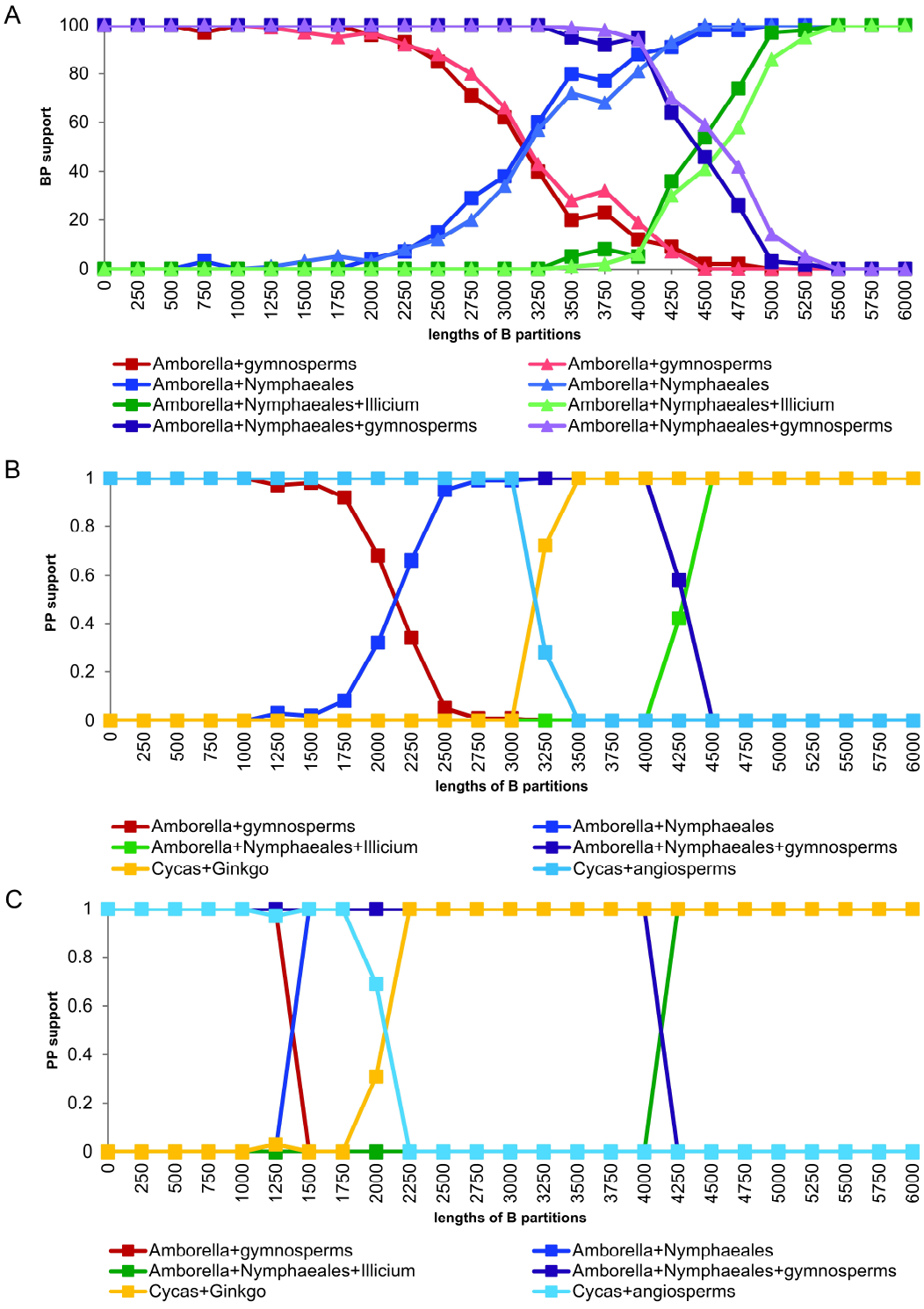

Fig. 3

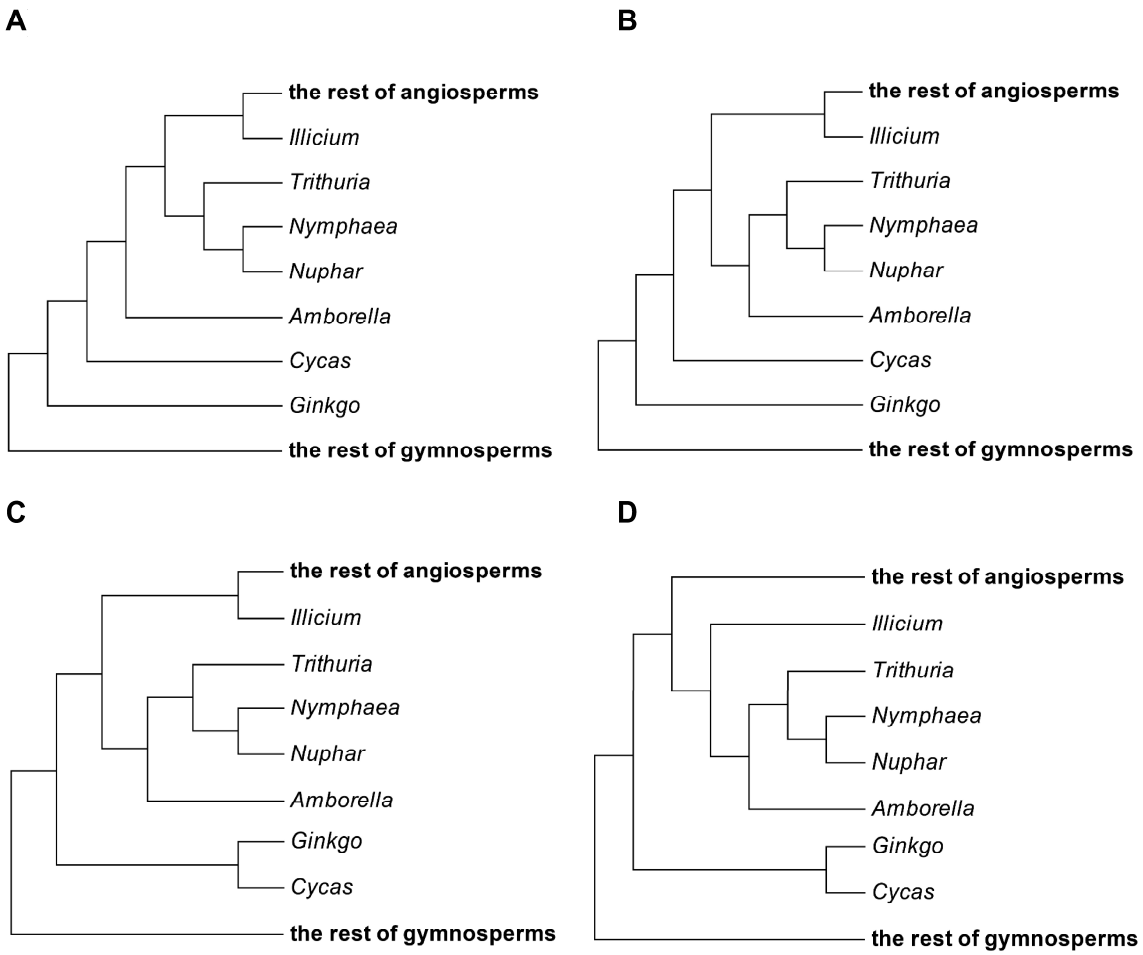

Fig. 4

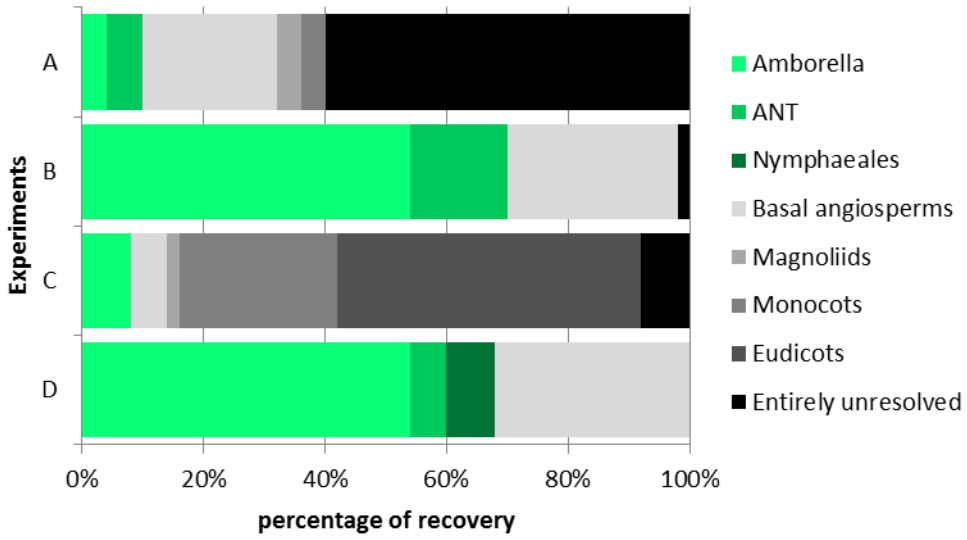

Fig. 5