Molecular Phylogenetics and Evolution 78 (2014) 334–348

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Dubious resolution and support from published sparse supermatrices: The importance of thorough tree searches

Mark P. Simmons [a,*], Pablo A. Goloboff [b,c]

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
[b] Consejo Nacional de Investigaciones Científicas y Técnicas, Miguel Lillo 205, 4000 S.M. de Tucumán, Argentina
[c] Instituto Miguel Lillo, Facultad de Ciencias Naturales, Miguel Lillo 205, 4000 S.M. de Tucumán, Argentina

ARTICLE INFO

Article history:
Received 9 April 2014
Revised 30 May 2014
Accepted 1 June 2014
Available online 11 June 2014

Keywords:
Maximum likelihood
Parsimony
Rapid bootstrap
RAxML
SH-like aLRT
Tree hybridization

ABSTRACT

We re-analyzed 10 sparse supermatrices wherein the original authors relied primarily or entirely upon maximum likelihood phylogenetic analyses implemented in RAxML and quantified branch support using the bootstrap. We compared the RAxML-based topologies and bootstrap values with both superficial- and relatively thorough-tree-search parsimony topologies and bootstrap values. We tested for clades that were resolved by RAxML but properly unsupported by checking if the SH-like aLRT equals zero and/or if the parsimony-optimized minimum branch length equals zero. Four of our conclusions are as follows. (1) Despite sampling nearly 50,000 characters, highly supported branches in a RAxML tree may be entirely unsupported because of missing data. (2) One should not rely entirely upon RAxML SH-like aLRT, RAxML bootstrap, or superficial parsimony bootstrap methods to rigorously quantify branch support for sparse supermatrices. (3) A fundamental factor that favors thorough parsimony analyses of sparse supermatrices is being able to distinguish between clades that are unequivocally supported by the data from those that are not; superficial likelihood analyses that quantify branch support using the bootstrap cannot be relied upon to always make this distinction. (4) The SH-like aLRT and parsimony-optimized-minimum-branch-length tests generally identify the same properly unsupported clades; the latter is a more severe test.

1. Introduction

For over 25 years molecular phylogeneticists have been generating sequence data for numerous species within most macroscopic eukaryotic lineages, and typically sample the same set of gene regions within each lineage (e.g., ITS, *matK*, *rbcL*, and *trnL-F* for vascular plants). This wealth of publicly available data, coupled with genomic studies that cover an increasingly diverse set of taxa, has enabled systematists to create supermatrices (Sanderson et al., 1998) that often contain upwards of 200 species and 10,000 characters without actually generating any novel sequence data. Because of their broad taxonomic reach, numerous species sampled, and the expectation that their inclusion of many thousands of characters will lead to accurate phylogenetic inference, these supermatrix studies are generally highly cited and referenced by numerous scientists outside of the systematics community. The taxonomic breadth and numbers of species and characters are impressive, but so is the enormity of tree space (Felsenstein, 1978a) and the percentage of inapplicable and missing data, which typically constitute the majority (and sometimes >95%; e.g., Peters et al., 2011) of the "sparse" supermatrices.

Missing data in empirical sparse supermatrices that consist largely or entirely of publicly available data are inevitably non-randomly distributed among species and gene regions. The potential for these non-randomly distributed missing data, in the context of other factors such as rate heterogeneity among characters and/or branches, to cause phylogenetic artifacts in maximum likelihood (Felsenstein, 1973) and/or Bayesian MCMC (Yang and Rannala, 1997) analyses has been forcefully argued as either a minor (e.g., Wiens and Morrill, 2011; Roure et al., 2013) or a major (e.g., Lemmon et al., 2009; Simmons 2012a,b; Dell'Ampio et al., 2014) problem in empirical studies.

Even without the non-randomly-distributed-missing-data problem, obtaining optimal trees for empirical matrices with hundreds or thousands of terminals is a difficult problem for which dedicated heuristic techniques have been developed because standard branch-swapping techniques (such as standard subtree pruning and regrafting) are likely to fail (e.g., Goloboff, 1999;

* Corresponding author. Address: Department of Biology, 200 West Lake Street, Colorado State University, Fort Collins, CO 80523-1878, USA. Fax: +1 970 491 0649.
E-mail address: psimmons@lamar.colostate.edu (M.P. Simmons).

http://dx.doi.org/10.1016/j.ympev.2014.06.002
1055-7903/© 2014 Elsevier Inc. All rights reserved.

Nixon, 1999; Roshan et al., 2004; Goloboff and Pol, 2007). Even after optimal trees have been identified, there is the problem of sufficiently sampling the breadth of all optimal trees so that systematic inferences are restricted to properly supported clades that are present in the strict consensus (Schuh and Polhemus, 1980; Nixon and Carpenter, 1996; Goloboff and Farris, 2001). Matrices with low phylogenetic signal, whether caused by inclusion of few parsimony-informative characters, character conflict, or the distribution of missing data (as in sparse supermatrices), are particularly liable to have multiple equally optimal trees (Maddison, 1991; Morrison, 2007; Sanderson et al., 2011), which makes accurate identification of the strict consensus especially important.

Given the expected difficulty of finding optimal trees and the reasons to expect multiple optima, it is curious that many prominent sparse-supermatrix studies (e.g., see Section 2.1 below) have relied exclusively upon likelihood analyses implemented in RAxML (Stamatakis, 2006) for phylogenetic inference. RAxML relies upon "lazy" and local subtree pruning and regrafting and only ever presents a single fully resolved optimal tree. Stamatakis et al. (2008, p. 770) asserted that rapid bootstrapping in RAxML "... solves—to a large extent—the computational problems associated with present-day full [maximum likelihood] analyses with a couple of hundred or a few thousand taxa." Peters et al. (2011, p. 10) were equally confident: "Unless one wants to analyze data sets that are significantly larger than ours [i.e., 1146 terminals and 88,626 characters], there is no computational or speed argument left to perform supertree or parsimony methods in favor of ML analyses." With respect to the issue of finding equally optimal trees, Bininda-Emonds and Stamatakis (2007) asserted that presenting a single fully resolved optimal tree is not problematic because the complexity of the likelihood (as opposed to parsimony) surface typically only allows for one or a few equally optimal trees. In contrast to Stamatakis et al. (2008), Siddall (2010) noted that in rapid bootstrapping the results are biased in favor of the original tree topology and Simmons and Norton (2014) showed how rapid bootstrapping with the GTRCAT model in RAxML can provide extremely high support values for simple 4-terminal polytomies and matrices that have no missing data. In contrast to Bininda-Emonds and Stamatakis (2007), Morrison (2007) argued that presenting a single fully resolved optimal tree in many cases constitutes specious precision that is not representative of the data.

Given the strong differences in opinion expressed by the above authors regarding the use of parametric methods to analyze sparse supermatrices as well as the suitability of lazy, local subtree-pruning-regrafting searches in RAxML to conduct those analyses, it is unclear whether the resulting trees should be embraced as "... presenting the state-of-the-art with respect to hypotheses of evolutionary relationships within the group" (Bininda-Emonds, 2011, p. 1; in reference to Peters et al. (2011)) wherein the bootstrap values are conservative estimates of branch support (Pyron and Wiens, 2011; Pyron et al., 2011) or rather an example from bioinformatics wherein, "Overzealous data mining is seen to have replaced carefully performed experimental analyses..." (Morrison, 2013, p. 349).

Two alternative (or perhaps complementary) approaches to test for properly unsupported clades in phylogenetic analyses wherein only a single optimal tree is presented (as in GARLI (Zwickl, 2006), PhyML (Guindon et al., 2010), and RAxML) are to check if the SH-like aLRT (Shimodaira–Haesgawa-like approximate likelihood ratio test; Anisimova and Gascuel, 2006; Guindon et al., 2010) value equals zero and to check if the parsimony-optimized minimum branch length equals zero (Simmons and Norton, 2014; Simmons and Randle, 2014). These two approaches have the advantage of requiring little additional computational power beyond the initial tree search and of being implemented in widely used programs. Therefore, they are readily applicable to supermatrices containing thousands of terminals. Both approaches are capable of identifying properly unsupported clades in simple simulated examples (4-terminal polytomies (Simmons and Norton, 2014); 8-terminal trees with various distributions of missing data or other ambiguous characters (Simmons and Randle, 2014)), but the question remains: how do they perform on large empirical sparse supermatrices? In such cases limiting SH-like aLRT comparisons to alternative topologies that are connected by nearest-neighbor-interchange swaps may grossly overestimate support when other swaps (e.g., subtree-pruning regrafting to a distant node) produce trees of the same likelihood. Identifying properly unsupported clades in sparse supermatrices is arguably the most important context for these two alternative approaches because of the high probability of having numerous properly unsupported clades given the superficial tree searches that are employed relative to the vast number of possible trees and the very high percentage of missing data in the matrix.

In this study we re-analyzed 10 published sparse supermatrices wherein the original authors relied primarily or entirely upon likelihood analyses implemented in RAxML and quantified branch support using the bootstrap. We compared the fully resolved RAxML-based topologies and bootstrap values with both superficial and relatively thorough-tree-search parsimony topologies (either fully resolved or the strict consensus) and bootstrap values. We also tested for properly unsupported clades on the RAxML topologies by checking if the SH-like aLRT value equals zero and checking if the parsimony-optimized minimum branch length equals zero. By making these comparisons among alternative tree-search methods and ways of quantifying branch support, we sought to quantify the extent to which these sparse supermatrices contain properly unsupported clades and inflated branch-support values based on the limitations of both superficial tree searches as well as the non-random distributions of missing data. We found unsupported resolution and inflated branch support in all 10 sparse supermatrices, though the extent to which these problems occurred varies widely.

2. Methods

2.1. Supermatrices sampled

The following 10 prominent recently published supermatrices were selected for inclusion in this study: Fabre et al. (2009; hereafter "Fabre"), Hedtke et al. (2013; hereafter "Hedtke"), Hinchliff and Roalson (2013; hereafter "Hinchliff"), Nyakatura and Bininda-Emonds (2012; hereafter "Bininda"), Peters et al. (2011; hereafter "Peters"), Pyron and Wiens (2011; hereafter "Wiens"), Pyron et al. (2011; hereafter "Pyron"), Soltis et al. (2013; hereafter "Soltis"), Springer et al. (2012; hereafter "Springer"), and van der Linde et al. (2010; hereafter "Linde"). These supermatrices include 180–2872 terminals, 5814–88,626 characters, and 66.7–98.4% missing or inapplicable data (Table 1). All of the matrices are based on sequence characters, all but one of these supermatrices include characters sampled from two or three genomes, and the taxa sampled range from Magnoliophyta (Hinchliff, Soltis) to Insecta (Hedtke, Linde), and Vertebrata (Bininda, Fabre, Peters, Pyron, Springer, Wiens; Table 1). Given the wide breadth of sampling with respect to numbers of terminals and characters, percent missing data or inapplicable entries, and genomes and taxa sampled, we hypothesize that our results will be broadly applicable to contemporary plant and animal sparse supermatrix studies in general.

In the three cases where the authors of the original studies analyzed two or more supermatrices, we selected the one that they focused on in their results and discussion (though Hedtke focused about equally on both of their matrices). For Hedtke we sampled

Table 1
Characteristics of supermatrices included in this study.

Matrix	# of terminals	# of characters	Percent missing or inapplicable	Genome(s) sampled	Ingroup taxon
Bininda	237	43,834	75.4	Mitochondrial, nuclear	Carnivora
Fabre	279	42,666	83.1	Mitochondrial, nuclear	Primates
Hedtke	1376	17,269	84.9	Nuclear only	Anthophila
Hinchliff	435	16,016	78.7	Mitochondrial, nuclear, plastid	Cyperaceae
Linde	180	14,912	73.6	Mitochondrial, nuclear	Drosophilidae
Peters	1146	88,626	98.4	Mitochondrial, nuclear	Hymenoptera
Pyron	767	5814	66.7	Mitochondrial, nuclear	Colubroidea
Soltis	950	48,465	94.8	Mitochondrial, nuclear, plastid	Saxifragales
Springer	372	61,199	68.6	Mitochondrial, nuclear	Primates
Wiens	2872	12,712	79.8	Mitochondrial, nuclear	Amphibia

the species-level analysis, for Hinchliff we sampled the fourth matrix (“Scaffold taxa only/rogues filtered”), and for Peters we sampled “subset 1.” The original matrices were variously e-mailed by the original authors (Bininda, Fabre) or downloaded from the author’s website (Peters), Dryad (Hinchliff, Soltis, Wiens) or TreeBASE (Hedtke, Linde, Pyron, Springer).

Seven of the 10 supermatrices were created entirely using sequences downloaded from GenBank. The exceptions were Linde, wherein they added “a limited amount of new sequence chosen to increase overlap” (p. 27); Pyron, wherein sequence data were added for two genes; and Springer, wherein sequence data were added for four genes.

The authors of the original studies employed a range of strategies to avoid problems with missing data. These include requiring sequences to be longer than a set minimum (Bininda, Wiens), eliminating characters for which most terminals have gaps (Peters), eliminating loci that were only sampled for a small percentage of terminals (Bininda, Fabre, Soltis), and ensuring that each cluster of putatively homologous sequences has taxonomic overlap with at least one other cluster (Soltis). Terminal-deletion strategies included eliminating all terminals that were only sampled for a single locus (Fabre, Hedtke, Peters, Springer), all terminals that were only sampled for a single locus with low variability (Pyron), all terminals that were not sampled for a particular locus (Peters) or loci (Hinchliff), terminals that were only sampled for relatively few nucleotides (Linde), species that were not sampled for at least one gene with other species in the same genus (Fabre), and terminals that were found to be resolved in disparate portions of alternative trees (Hinchliff).

Nine of the 10 studies relied entirely upon RAxML likelihood analyses for phylogenetic inference from their supermatrix or supermatrices (Table 2). For the RAxML analyses, the GTRCAT and/or GTRGAMMA models were used (together with PROTCAT for Peters), typically on partitioned data. Branch support was generally calculated using the rapid bootstrap with 100–1000 replicates (Table 2).

Bayesian MCMC analyses have been recognized as problematic to apply to matrices with numerous terminals because of difficulty obtaining stationarity and/or convergence between independent runs (Soltis et al., 2007; Marshall, 2010; Pyron and Wiens, 2011). Of the 10 studies sampled, only Linde performed parsimony (and Bayesian) phylogenetic analyses of their supermatrix in addition to likelihood. Fabre, Hedtke, Hinchliff, Pyron, Soltis, and Springer made no mention of parsimony-based phylogenetic analyses at all, while Bininda, Peters, and Wiens explicitly dismissed parsimony in favor of their RAxML-based likelihood analyses.

2.2. Modifications to RAxML results presented by the original authors

In six cases we used the RAxML optimal tree with bootstrap values mapped on it directly from the original studies. The four exceptions are as follows.

Fabre did not present actual bootstrap values in their Figs. 3–7, but rather used four symbols to indicate ranges of possible values. P.-H. Fabre (pers. comm. 2013) was unable to locate the RAxML output file with partitioned-analysis bootstrap values. Therefore we re-ran the bootstrap analyses in RAxML ver. 7.7.2 using rapid bootstrapping (-f a). To increase precision of the bootstrap values we ran 1000 pseudoreplicates instead of Fabre et al.’s (2009) 100 pseudoreplicates. RAxML ver. 7.7.2 does not implement the GTRMIX model that Fabre et al. (2009) applied in RAxML 7.0.4. Instead the bootstrap analyses were run using GTRCAT together with Fabre et al.’s (2009) partitions and then plotted on Fabre et al.’s (2009) optimal tree from their partitioned optimal-tree-search analysis.

Hinchliff presented majority-rule consensus (Margush and McMorris, 1981) bootstrap trees from their 300 pseudoreplicates rather than presenting the optimal RAxML tree with bootstrap values mapped onto it, as was done in the other studies. Hinchliff’s approach can be problematic because of undersampling-within-replicates artifacts, which RAxML is particularly susceptible to given that it only ever presents a single fully resolved tree

Table 2
Phylogenetic analyses applied to supermatrices applied by original authors.

Matrix	Program(s) applied	Model(s) applied in RAxML	Partition(s) applied	Bootstrap calculations	Bootstrap replicates
Bininda	RAxML 7.0.4	GTRMIX	74	Rapid bootstrap	1000
Fabre	RAxML 7.0.4	GTRMIX	1, 27	Standard bootstrap	100
Hedtke	RAxML 7.2.8	GTRCAT	60[a]	Standard bootstrap	100
Hinchliff	RAxML 7.2.6	GTRCAT	18	Rapid bootstrap	300
Linde	GARLI, MrBayes, PAUP*, RAxML 7.0.4	GTRGAMMA	1, 9 or 15	Rapid bootstrap	250
Peters	RAxML 7.2.8	GTRCAT + PROTCAT	32	Rapid bootstrap	560
Pyron	RAxML 7.0.4	GTRGAMMA	3	Rapid bootstrap	1,000
Soltis	RAxML-VI-Light 1.0.5	GTRCAT	1	Standard bootstrap	200
Springer	RAxML 7.2.8	GTRGAMMA, GTRCAT	79	Rapid bootstrap	500
Wiens	RAxML 7.0.4	GTRGAMMA	32	Rapid bootstrap	100

[a] Hedtke examined six different partitioning schemes and ultimately applied the most complex one (S. Hedtke, pers. comm., 2013).

(Goloboff and Farris, 2001; Simmons and Freudenstein, 2011; Simmons and Goloboff, 2013). To make Hinchliff's trees and bootstrap values more directly comparable to those from the other nine supermatrices sampled, we performed 100 optimal-tree searches in RAxML ver. 7.7.2 using the GTRCAT model. To obtain bootstrap values for all branches of the fully resolved optimal tree found we then re-ran the bootstrap analysis using 300 pseudoreplicates, the rapid bootstrap, and the GTRCAT model.

Linde focused on their partitioned RAxML analysis in their paper, presenting it as their Fig. 3. But the character partitions cited in their Table 1 do not mention the *per* intron or match those in their TreeBase submission (study 10691). S.T. Steppan (pers. comm. 2013) was unable to locate the original RAxML files. Therefore we-ran searches for the maximum likelihood tree (100 replicates) as well as the bootstrap analyses in RAxML ver. 7.7.2 with the 15 partitions in the TreeBase file using the GTRGAMMA model and rapid bootstrapping (-f a) as suggested by S.T. Steppan. We ran 1000 pseudoreplicates.

Soltis presented bootstrap values in their Appendix S1b, but the tree includes numerous polytomies because they collapsed branches in the optimal RAxML tree that r8s (Sanderson, 2003) estimated to be of zero length. Therefore, SumTrees ver. 3.3.1 (Sukumaran and Holder, 2010) was used to map the bootstrap values from the 200 bootstrap trees onto the single fully resolved optimal tree, both of which Soltis posted in Dryad (doi:10.5061/dryad.h4070).

2.3. Parsimony tree searches in TNT

Two-stage tree searches for the most parsimonious trees for equally weighted characters were performed in the TNT ver. 1.1 (Goloboff et al., 2008) that was released in January 2013. Tree building was performed in the first stage and tree hybridization was performed in the second stage. The series of commands used for both stages, including descriptions of the functions of non-standard commands used, are posted as supplemental online data at: http://rydberg.biology.colostate.edu/Research/.

TNT's macrolanguage was used for the tree-hybridization phase of tree searches, and use of this language requires that matrices be uploaded in TNT format. Conversion of the matrices into TNT format was generally straightforward and just involved text editing and/or use of PAUP* ver. 4.0b10 (Swofford, 2001) to remove unnecessary information. The exception was Peters' subset-1 matrix, which was converted into TNT format by using Mesquite ver. 2.75 (Maddison and Maddison, 2013) to convert the 17,562 nucleotide characters and Geneious ver. 6.1.0 (Kearse et al., 2012) to convert the 71,064 amino-acid characters, followed by concatenation of the two sub-matrices.

One thousand replicates of tree building were performed using random non-repeating integers between 1 and 32,767 (the effective maximum in the January 2013 release of TNT) generated by http://www.random.org/integer-sets/. Tree building incorporated tree drifting, tree fusing, and sectorial searches (Goloboff, 1999). Each replicate of tree building was initiated by performing 100 random-addition-sequence (RAS) searches that each only held a single tree. Sectorial searches were then performed beginning with 32 exclusive sectors and progressively working down to 10 exclusive sectors in decrements of two sectors at a time. Global tree-bisection-reconnection (TBR) searches were performed after each round of sectorial searches. Within each exclusive sector of terminals, 10 iterations of RAS + TBR were performed. Five cycles of tree drifting were performed in sectors of $\geqslant 50$ terminals; five rounds of tree fusing were performed in sectors of <50 terminals. Within each of the five cycles of tree drifting, trees were accepted up to two steps longer with a probability that is partially determined by a relative fit difference of 0.25.

A second set of sectorial searches was then performed within each replicate of tree building with eight exclusive sectors and progressively working down to two exclusive sectors in decrements of one sector at a time. Sectors with widely disparate numbers of terminals were allowed and global TBR was performed after each round of sectorial searches. In addition to the commands used above for 32–10 exclusive sectors, each of the 8–2 sectors was subdivided five times into five exclusive sub-sectors each time followed by TBR within each sector. Ten replicates of RAS + TBR with five cycles of tree drifting were then performed. Each of the ten trees found for each sector was then fused with the original topology of the sector and only the best trees were retained. Finally, all of the most parsimonious trees found were saved as fully bifurcating trees.

All most parsimonious trees found for 100 replicates of tree building were concatenated into a single file. Each of the 10 resulting files was uploaded into an independent round of tree-hybridization with up to 500,000 most parsimonious trees held (except for the Soltis matrix for which up to 400,000 trees were held and the Hedtke and Peters matrices for which up to 100,000 trees were held because of RAM limitations). Ten random non-repeating integers between 1 and 32,767 were used as seeds. Tree hybridization was performed using TNT's macrolanguage with two user variables: the first tree in memory and the next tree in memory. A loop was created wherein exchanges of groups were made from the first tree held in memory to each of the other trees (sequentially); the groups exchanged contain no more than 10 unshared terminals. This process was repeated until exchanges were made with all other trees. After that the process was repeated using the second tree in memory, then the third tree, etc. After two trees were hybridized only the shortest tree was retained for subsequent hybridization with other trees. Once the loop was completed only the shortest trees were retained.

An exclusive sectorial search was then performed on the shortest trees retained from tree-hybridization using the same settings as those used in the first set of sectorial searches from tree-building stage except the sector number ranged from 32 to 2 in decrements of one. Only the most parsimonious trees found were retained in memory, after which TBR-collapsing (Goloboff and Farris, 2001) was applied and the most parsimonious and strict consensus trees were saved. The lengths of the most parsimonious trees found in each of the ten independent rounds of tree-hybridization were determined. A final strict consensus was then calculated from those rounds that produced most parsimonious trees.

Bootstrap frequencies, rather than jackknife (Farris et al., 1996) or GC values (Goloboff et al., 2003) were generated using TNT so as to make the parsimony-based support values more directly comparable to the likelihood-based support values generated by the original authors of the 10 supermatrices sampled (Table 2). One thousand bootstrap pseudoreplicates were performed using a second set of non-repeating integers between 1 and 32,767. Because the 32-bit versions of TNT released prior to September 2013 do not properly bootstrap characters in matrices with >32,767 characters (they only re-sample the first 32,767 characters), bootstrap analyses of the Bininda, Fabre, Soltis, and Springer matrices were limited to the variable characters, and bootstrap analyses of the Peters matrix were limited to the parsimony-informative characters. This limitation of TNT was fixed in the September 2013 release.

Each bootstrap pseudoreplicate held up to 100,000 most parsimonious trees. Similar searches were performed as in the tree-building phase described above; tree-hybridization was not performed. Each pseudoreplicate consisted of 1000 RAS searches that each only held a single tree, followed by exclusive sectorial searches with 32–2 sectors in decrements of one. The rest of the commands were identical to those used in the first set of sectorial searches described above. TBR-collapsing was applied.

In addition to the relatively thorough heuristic tree searches described above, a second set of TNT analyses were performed to roughly emulate superficial RAxML tree searches, albeit in the context of parsimony rather than likelihood. One hundred tree searches were performed using time to select seeds for the RAS searches that applied subtree-pruning-regrafting branch swapping with only a single tree held per search. Fully resolved trees were retained; zero-length branches were not collapsed. After completion of the search only the first tree saved by TNT was considered; any others, when applicable, were deleted. One thousand bootstrap pseudoreplicates were performed. Each bootstrap pseudoreplicate included a single RAS + subtree-pruning-regrafting search and a single fully resolved most parsimonious tree was saved.

Bootstrap values were mapped onto the strict consensus (for the relatively thorough searches) or the first most parsimonious tree saved (for the superficial searches) using SumTrees. To import the TNT trees into SumTrees they were converted into NEXUS format using MacClade ver. 4.08 (Maddison and Maddison, 2001) or Mesquite and then saved without a translation table by PAUP*. The 20 SumTrees output trees are posted as supplemental online data at: http://rydberg.biology.colostate.edu/Research/.

2.4. Likelihood analyses in RAxML

SH-like aLRT branch-support values were calculated by RAxML ver. 7.7.2 using the model(s) and partition(s) applied by the original authors as presented in Table 2 with the following qualifications and clarifications. Bininda and Fabre used the GTRMIX model, which is not implemented in RAxML ver. 7.7.2; we used the GTRGAMMA model instead. Springer used the GTRGAMMA model to search for the optimal trees and the GTRCAT model for the bootstrap pseudoreplicates; we used the GTRGAMMA model for the SH-like aLRT. For Fabre we ran the analysis with 27 partitions and for Linde we ran the analysis with 15 partitions. SH-like aLRT values were generated using the "-f J" command after uploading the optimal RAxML tree reported by the original authors (with the qualifications noted in Section 2.2 above). The RAxML output trees with SH-like aLRT values mapped on them are posted as supplemental online data at: http://rydberg.biology.colostate.edu/Research/.

Sanderson and Kim (2000) and Xia (2006) noted that because of computational shortcuts implemented in heuristic likelihood tree searches it might be possible for more thorough tree-search methods implemented using a different optimality criterion, such as parsimony, to actually find more likely trees than the likelihood tree searches. We tested whether our relatively thorough TNT-based parsimony searches produced more likely trees than did RAxML. Although possible, this situation seems unlikely given the complexity of the GTR Q-matrices, rate heterogeneity among sites, and partitioning schemes applied in the RAxML analyses (as opposed to, e.g., likelihood analyses using the Jukes and Cantor (1969) model without rate heterogeneity).

Different likelihood tree-search programs can report different likelihood values for the same tree (Morrison, 2007; Sundberg et al., 2008). Therefore we calculated the likelihoods of fully bifurcating parsimony trees (that were found after tree-hybridization) using RAxML ver. 7.7.2 with the "-f e" command given that RAxML was used to find the most likely trees reported by the original authors for the ten matrices (Table 2). Within RAxML the same models and partitions were applied to the parsimony-based trees as those described above for calculating SH-like aLRT values. Identical trees found between the 10 separate TNT tree-hybridization analyses were eliminated. When >10 trees remained then 10 of them were selected using a set of non-repeating random integers between 1 and the total number of unique bifurcating most parsimonious trees saved.

To test for significant differences in likelihoods between the RAxML and TNT trees we implemented the Shimodaira–Hasegawa test (SH test; Shimodaira and Hasegawa, 1999) in RAxML using the "-f H" command such that model parameters are re-estimated for each tree tested. The RAxML ver. 8.0.X manual (Stamatakis, 2014, p. 7) explicitly states, "Warning: never compare alternative tree topologies using their CAT-based likelihood scores!" But A. Stamatakis (pers. comm. 2014) noted that the "-f H" command may be used with the GTRCAT model "because the rate-to-site assignments are identical for all trees under CAT when using -f H." With respect to likelihoods of the trees found using the GTRCAT model presented below, they include a final GTRGAMMA optimization using the RAxML default settings.

2.5. Parsimony branch lengths

MacClade was used to identify branches in the RAxML optimal trees presented by the original authors (with the qualifications noted in Section 2.2 above) that can be optimized by Fitch (1971) parsimony to have a length of zero. That is, parsimony-based optimization was applied to the likelihood-based topologies. Because MacClade can only load 32,000 characters, only variable characters from the Bininda, Fabre, Soltis, and Springer matrices were uploaded into MacClade after excluding constant characters in PAUP*. Because the Peters matrix contains 39,148 variable characters, only parsimony-informative characters were uploaded from the Peters matrix. Parsimony-uninformative characters cannot raise the minimum possible internal branch length above zero so our reliance on the parsimony-informative characters should not be determinate to the results.

MacClade was unable to load the Wiens matrix with 2872 terminals. To bypass this limitation we separated two well supported clades (Hyloidea, 98% BS; Plethodontidae, 98% BS) from the rest of the terminals in the matrix. We then used TreeGraph2 ver. 2.0.45-197 (Stöver and Müller, 2010) to subdivide these clades from the rest of the terminals in the tree file. Each of the three sub-matrices and their corresponding sub-trees were then uploaded into MacClade independently of each other.

Parsimony branch lengths at the root of a tree are by definition ambiguously optimized because each character-state change can be optimized onto either of the two branches. Hence these two branches were not counted as having a minimum optimized branch length of zero.

2.6. Quantification of branch support

Scaled support was quantified by using Simmons and Webb's (2006) suggested four cutoffs (rounded to integers) of 63%, 86%, 95%, and 98% bootstrap support, which scaled bootstrap support to that provided by 1–4 uncontradicted synapomorphies. 63–85% bootstrap was scaled to 0.25, 86–94% bootstrap was scaled to 0.5, 95–97% bootstrap was scaled to 0.75 and ⩾98% bootstrap was scaled to 1. Clades with 50–62% bootstrap were also incorporated by scaling them to 0.125. Scaled support was then summed across all internal branches on a single fully resolved optimal tree identified (for superficial parsimony and RAxML) or the strict consensus (for relatively thorough parsimony). Scaled support is directly applicable to bootstrap and jackknife values but the cutoffs are expected to be arbitrary when applied to SH-like aLRT values. Nonetheless, scaled support was extended to SH-like aLRT values so as to provide a rough comparison with the scaled-support values that are based on bootstrap values.

3. Results

3.1. Optimal trees

Results from each of the two stages of the relatively thorough parsimony tree searches are presented for the 10 supermatrices in Table 3. Aside from Linde, all 1000 tree-building replicates identified the shortest known trees for the matrices with <400 terminals (Bininda, Fabre, and Springer; Table 1). In contrast, 0–0.6% of the trees held from the 1000 tree-building replicates are the shortest known for most of the matrices with >700 terminals (Peters, Pyron, Soltis, Wiens; not including Hedtke; Table 1). Tree hybridization only found shorter trees than did tree building for two of the matrices (Peters and Soltis), and only in one of the 10 separate tree-hybridization analyses (Table 3).

The resolution, summed scaled support, and average percent support from both sets of parsimony analyses (superficial and relatively thorough) and both sets of RAxML analyses (including both bootstrap values and SH-like aLRT values) are presented in Table 4. The numbers of clades in the optimal trees for the RAxML and superficial-parsimony analyses are by definition the maximum possible. By contrast, the relatively thorough parsimony strict consensus trees had 7.6% (Fabre) to 100% (Soltis) fewer clades resolved—with the strict consensus for Soltis being entirely unresolved. The next most extreme case was for Hinchliff, with 42.6% fewer clades resolved on the strict consensus.

The Soltis parsimony strict consensus is completely unresolved for the matrix of 950 terminals (Table 4). To help identify problematic taxa that are behaving as wildcards (Nixon and Wheeler, 1991) the matrix was compartmentalized (Maddison et al., 1984) into the following six submatrices based on the RAxML topology presented by Soltis in their supplemental appendix 1A: (1) Crassulaceae (monophyletic, 365 terminals with 2099 parsimony-informative [PI] characters), (2) Halloragaceae alliance (monophyletic, 94 terminals with 1126 PI characters), (3) outgroups (paraphyletic, 45 terminals with 929 PI characters), (4) Grossulariaceae + Iteaceae + Pterostemonaceae (paraphyletic, 102 terminals with 681 PI characters), (5) Saxifragaceae (monophyletic, 243 terminals with 2029 PI characters), and (6) the woody clade (monophyletic, 101 terminals with 1776 PI characters).

Parsimony-based TNT optimal-tree searches were then performed on these six submatrices using the same procedure as applied for the matrix as a whole. All 10 sets of tree-hybridization analyses identified optimal trees of the same number of steps for each of the six submatrices. The Crassulaceae and Saxifragaceae strict-consensus trees are completely unresolved, the outgroup tree contains 13 clades, the Grossulariaceae + Iteaceae + Pterostemonaceae tree contains 37 clades, the woody-clade tree contains 63 clades, and the Halloragaceae alliance contains 81 clades. Hence for only two of the six submatrices are the majority of the possible clades resolved in the strict consensus, which indicates that the problem of properly unsupported resolution is widespread.

The Crassulaceae and Saxifragaceae submatrices of only parsimony-informative characters were then manually examined in MacClade to identify wildcard terminals, for which several examples follow. In the Crassulaceae submatrix, *Sedum humifusum* Rose is not sampled for any of the 2099 parsimony-informative characters (Simmons, 2014). The characters it is sampled for are also scored for only 15 or 16 other terminals in the complete 950-terminal matrix, and of these only two (*Crassula marnieriana* H. Huber & H. Jacobsen and *Kalanchoe pinnata* Pers.) are within the Crassulaceae. These three Crassulaceae species are not resolved as a three-terminal clade in the RAxML tree. If they were then the resolution of *Sedum humifusum* as a member of a nested two-terminal clade might have been unequivocally supported given that all three terminals are scored for common characters (Malia et al., 2003; Simmons, 2014). Rather, *Sedum humifusum* is separated by at least 10 branches from the other two species.

The "chkmoves;" command was applied in TNT after uploading the entire Soltis matrix and the single most parsimonious tree that we retained after tree-hybridization. The "chkmoves" command may be used to identify wildcard terminals in seconds based on their number of possible moves, maximum distance (based on number of branches), or re-rooting depth among equally parsimonious trees (http://tnt.insectmuseum.org/index.php/Commands/chkmoves). Ten species were identified as having more than 100 possible moves, including *Sedum humifusum*, with 674 possible moves.

Some other members of Crassulaceae with high percentages of missing data among the parsimony-informative characters in the Crassulaceae submatrix are as follows. *Crassula perforata* Thunb. (96.5% inapplicable/missing data) is only sampled for one block of parsimony informative characters that are also scored for five other Crassulaceae terminals. Seven *Crassula* species (*C. colligata* Toelken, *C. manaia* A.P. Druce & Sykes, *C. mataikona* A.P. Druce, *C. moschata* G. Forst., *C. perfoliata* L., *C. pseudohemisphaerica* Friedrich, and *C. sieberiana* (Schult. & Schult.f.) Druce) with 92.8–92.9% inapplicable/missing data are only scored for one block of parsimony informative characters that are also scored for eight other Crassulaceae terminals.

In the Saxifragaceae submatrix, *Saxifraga stolonifera* Curtis (96.2% inapplicable/missing data) is only sampled for one block of parsimony informative characters that are also scored for three other Saxifragaceae terminals. *Saxifraga retusa* Gouan (97.8% inapplicable/missing data) is only sampled for one block of parsimony informative characters that are also scored for 10 other Saxifragaceae terminals. *Boykinia intermedia* (A. Heller) G.N. Jones (99.3% inapplicable/missing data) is only sampled for one block of parsimony informative characters that are also scored for 12 other Saxifragaceae terminals. *Saxifraga hirsuta* L. (97.5% inapplicable/missing

Table 3
Parsimony-tree-search results from each of the two-stage searches.

Matrix	Tree building		Tree hybridization	
	Minimum length	% Of all trees held	Minimum length	# Of the 10 separate analyses
Bininda	106,678	100	Same	10
Fabre	79,726	100	Same	10
Hedtke	91,540	73.1	Same	10
Hinchliff	27,092	99.9	Same	10
Linde	35,890	97.8	Same	10
Peters	100,874	0.1	100,871	1
Pyron	95,038	0.4	Same	4
Soltis	36,805	0.1	36,801	1
Springer	111,402	100	Same	10
Wiens	412,008	0.6	Same	5

Table 4
Resolution, summed scaled support, and average percent support from the different phylogenetic analyses. Summed scaled support and percentages are rounded to the nearest integer.

Matrix	# Clades in optimal tree/consensus			Summed scaled support				Average percent support			
	RAxML	Superficial parsimony	Thorough parsimony	RAxML bootstrap	RAxML SH-like aLRT	Superficial parsimony	Thorough parsimony	RAxML bootstrap	RAxML SH-like aLRT	Superficial parsimony	Thorough parsimony
Bininda	234	234	194	145	172	119	86	82	90	76	74
Fabre	276	276	255	196	206	155	158	88	89	83	84
Hedtke	1373	1373	1000	640	735	495	474	71	77	63	74
Hinchliff	432	432	248	190	234	110	75	72	78	56	63
Linde	177	177	149	88	114	53	51	75	85	58	63
Peters	1143	1143	1045	258	538	196	189	50	78	41	44
Pyron	764	764	627	339	436	239	240	70	82	55	64
Soltis	947	947	0	292	452	128	0	56	72	39	N/A
Springer	369	369	321	253	272	184	179	86	89	78	80
Wiens	2869	2869	2512	1455	1835	1137	1089	75	85	66	70

Table 5
Pairwise differences in average percentages between the four methods of quantifying branch support for the subset of clades wherein both methods provided positive support values on the RAxML reference-tree topology. Positive values indicate that the first method listed provided higher average support whereas negative values indicate that the second method listed provided higher average support.

Matrix	RAxML bootstrap vs.			SH-like aLRT vs.		Superficial parsimony vs.
	SH-like aLRT	Superficial parsimony	Thorough parsimony	Superficial parsimony	Thorough parsimony	Thorough parsimony
Bininda	−6.2	7.0	11.4	9.6	14.4	3.6
Fabre	−1.8	5.0	5.6	6.6	7.3	1.3
Hedtke	−9.9	6.1	6.0	11.6	10.8	0.9
Hinchliff	−11.0	11.8	16.9	17.9	21.6	6.5
Linde	−10.7	10.2	8.7	17.3	15.7	−0.8
Peters	−30.0	8.8	9.6	25.5	25.6	0.7
Pyron	−12.0	11.1	9.2	15.6	13.1	−0.6
Soltis	−20.5	9.8	N/A	23.7	N/A	N/A
Springer	−3.1	7.0	7.9	7.9	8.5	1.7
Wiens	−10.3	6.3	6.2	12.5	12.1	0.4

data) is only sampled for one block of parsimony informative characters that are also scored for 19 other Saxifragaceae terminals.

3.2. Branch-support values

For all 10 matrices the ranking for summed scaled support was RAxML SH-like aLRT > RAxML bootstrap > superficial parsimony, and in eight of the 10 matrices superficial parsimony > relatively thorough parsimony (Table 4). The differences in summed scaled support between the four methods ranged from minor to dramatic. RAxML bootstrap values provided 4.9% (Fabre) to 52.0% (Peters) lower summed scaled support than did the RAxML SH-like aLRT, while the superficial parsimony bootstrap provided 17.9% (Bininda) to 56.2% (Soltis) lower summed scaled support than did the RAxML bootstrap. Finally, the relatively thorough parsimony bootstrap provided 1.9% higher (Fabre) to 100% lower summed scaled support than did the superficial parsimony bootstrap.

Similar differences were observed when examining average percent support values, with the exception that the relatively thorough parsimony bootstrap analyses generally provided higher support for individual clades than did the superficial bootstrap analyses (Table 4). This apparent discrepancy with the scaled-support values is largely caused by average-percent-support values being calculated only for those clades resolved in the strict consensus for the relatively thorough parsimony analyses, whereas it was calculated for every clade in the fully resolved superficial tree. Note that average-percent support includes values <50%. Pairwise comparisons of RAxML bootstrap values for all clades in the RAxML-bootstrap reference tree and, alternatively SH-like aLRT, superficial parsimony bootstrap, and relatively thorough parsimony bootstrap values are presented for each of the 10 matrices in Figs. S1–S5.

To make comparisons between percent support values directly comparable at the individual-clade (rather than entire tree) level, pairwise differences in average percentages between the four methods of quantifying branch support for the subset of clades wherein both methods provided positive support values for the reference-tree topology are presented in Table 5. The numbers of pairwise comparisons used to calculate pairwise differences in average percentages between the four methods of quantifying branch support for the subset of clades wherein both methods provided positive support values for the reference-tree topology are presented in Table S1.

The same general patterns observed for differences in summed scaled support (Table 4) are also evident when making pairwise comparisons at the individual-clade level, albeit with lower extremes (Table 5). RAxML bootstrap values are 1.8–30.0% lower than SH-like aLRT values but 5.0–11.8% and 5.6–16.9% higher than superficial and relatively thorough parsimony bootstrap values, respectively. In comparison, the differences between superficial and relatively thorough parsimony bootstrap values for individual clades were relatively minor (6.5% higher to 0.8% lower).

3.3. Topological conflict between RAxML analyses

Calculation of SH-like aLRT values in RAxML used the optimal RAxML topology upon which bootstrap values were mapped, but the output tree differed for 0.5–4.7% of the branches for the uploaded tree (Table 6). Support values for the conflicting clades were generally low (averaging 22–56% bootstrap and 35–66% SH-like aLRT), but could be extreme (99% bootstrap vs. 90% SH-like aLRT for a pair of conflicting clades from Peters). These presumably represent instances where the uploaded RAxML tree is not nearest-neighbor-interchange optimal (Stamatakis, 2014).

Table 6
Cases of topological conflict between the RAxML optimal tree with bootstrap percentages mapped onto it and the RAxML optimal tree reported with SH-like aLRT percentages mapped onto it. Average bootstrap and SH-like aLRT percentages are rounded to the nearest integer.

Matrix	Clades		Bootstrap (%)		SH-like aLRT (%)		Maximum disparity in %
	Number	Percent	Average	Maximum	Average	Maximum	
Bininda	2	0.9	56	99	65	72	171
Fabre	11	4.0	40	66	28	73	105
Hedtke	15	1.1	25	85	35	71	95
Hinchliff	8	1.9	26	53	61	91	123
Linde	2	1.1	37	50	39	68	92
Peters	19	1.7	20	99	37	90	174
Pyron	36	4.7	28	60	46	89	125
Soltis	18	1.9	22	42	66	100	132
Springer	2	0.5	32	33	41	51	82
Wiens	84	2.9	28	85	37	86	117

3.4. Topological congruence between likelihood and parsimony

Topological congruence (i.e., taxonomic congruence) between RAxML clades with ⩾50% bootstrap support and parsimony trees is presented in Table 7 for each of the 10 studies. Congruence was quantified in four ways; for each way the reference clades were those RAxML clades with ⩾50% bootstrap, so as to eliminate clades that have very weak support and to limit comparisons to clades that might be discussed in empirical studies. The first approach was to quantify the overall success of resolution for the topology only (i.e., number of identical clades minus the number of conflicting clades; Simmons and Webb, 2006), while the second approach was to incorporate scaled support in the manner described in Section 2.6 above. The third approach was to only record conflicting clades and the fourth approach was to incorporate scaled support for the conflicting clades. The RAxML- and both parsimony-derived topologies were generally congruent based on all four approaches.

Topological congruence for the relatively thorough parsimony analyses was limited to those clades resolved on the strict consensus whereas the superficial parsimony analyses were based on fully bifurcating trees. This is the reason for the large discrepancies between the parsimony analyses for the Soltis matrix (Table 7) wherein the strict consensus is entirely unresolved (Table 4).

Setting aside the Soltis matrix, the overall success of resolution was greater for the superficial parsimony analyses for six matrices, identical to that for the relatively thorough parsimony analyses for one matrix, and lower for two matrices (Table 7). Likewise for the averaged overall success of resolution. In contrast, the relatively thorough parsimony analyses had fewer conflicting clades for all nine matrices. Hence the greater overall success of resolution for the superficial parsimony analyses is based on their having more clades that are congruent with the RAxML clades with ⩾50% bootstrap support rather than fewer conflicting clades. This strongly suggests that these congruent clades resolved by RAxML are also artifacts of superficial tree searches. The conflicting clades were typically very weakly supported for both parsimony analyses, which is evident when comparing the resolution-only vs. averaged conflicting results.

3.5. SH-like aLRT = 0 vs. parsimony minimum branch length = 0

The number of branches on the most likely reported RAxML trees that were found to have a parsimony minimum branch length of zero ranged from 5 to 188 (average 59; Table 8). Similarly, the number of branches with SH-like aLRT support of zero ranged from 2 to 152 (average 43). Both of these methods were in agreement that the clades in question are unsupported in many, but by no means all, cases (average 33).

With respect to the remaining cases, clades with a parsimony minimum branch length of zero frequently (average 19) received ⩾50% SH-like aLRT support (Table 8). On the other hand, the remaining clades with SH-like aLRT support of zero generally (average 6) were not resolved on the parsimony strict consensus, and of those that were, they were generally (average 3) contradicted in the parsimony tree and received ⩽50% bootstrap support. So, there is an asymmetric relationship between the two methods when they disagree, with the parsimony method indicating that many clades which SH-like aLRT considers supported are actually unsupported.

For the 189 clades wherein the parsimony minimum branch length was zero but the SH-like aLRT support was ⩾50%, we examined the data matrix to try and identify the cause of the discrepancy and, by extension, whether the parsimony or the SH-like aLRT result should be used to determine whether these clades are properly unsupported. Fig. 1 presents simplified contrived examples of the scored-character and character-state distributions observed in the empirical matrices. The scored-character distributions shown in A and B would be assigned to Simmons and Randle's (2014) third hypothesis, C would be assigned to their

Table 7
Congruence between RAxML clades with ⩾50% bootstrap support and parsimony trees. Parsimony results are presented as: superficial searches | relatively thorough searches. Averaged overall success of resolution and averaged conflicts only are rounded to the nearest integer.

Matrix	# RAxML clades		Overall success of resolution		Conflicts only	
	All	⩾50% BS	Resolution only	Averaged	Resolution only	Averaged
Bininda	234	205	153|128	109|75	−26|−21	−6|−4
Fabre	276	254	216|212	148|153	−19|−12	−2|−1
Hedtke	1373	1010	736|739	473|459	−137|−58	−7|−4
Hinchliff	432	332	228|187	108|75	−52|−17	0|0
Linde	177	141	83|81	51|50	−29|−21	0|0
Peters	1143	546	399|415	185|180	−73|−55	−1|−1
Pyron	764	550	352|352	229|230	−98|−74	−3|−2
Soltis	947	546	316|0	122|0	−78|0	−2|0
Springer	369	337	243|240	173|173	−47|−29	−5|−2
Wiens	2869	2254	1659|1639	1086|1038	−296|−205	−13|−14

Table 8
Comparison between SH-like aLRT percentages and Fitch-parsimony-optimized branches with a minimum length of zero.

Matrix	Minimum 0-length	SH-like aLRT = 0	Full agreement (0 + 0)	Minimum 0-length but SH-like aLRT ⩾50%	Minimum 0-length but SH-like aLRT ⩽−50%	SH-like aLRT = 0; minimum length >0 but parsimony consensus unresolved	SH-like aLRT = 0; minimum length >0 but clade contradicted by parsimony (generally <50%)	SH-like aLRT = 0; minimum length >0 and clade resolved on parsimony consensus with ⩾50% bootstrap[a]
Bininda	5	3	3	2	0	0	0	0
Fabre	10	3	3	7	0	0	0	0
Hedtke	139	127	105	31	3	17	3	2
Hinchliff	61	39	32	26	3	4	2	1 (48%)
Linde	9	2	2	6	1	0	0	0
Peters	44	50	32	10	2	2	12	4
Pyron	7	13	2	4	1	4	6	1 (32%)
Soltis	188	152	131	51	6	21	0	0
Springer	13	2	1	12	0	1	0	0
Wiens	58	38	15	40	3	12	7	4 (44%)

[a] Unless otherwise noted.

fourth hypothesis, E would be assigned to their first or second hypothesis, and F would be assigned to their eighth hypothesis. None of the six scored-character and character-state distributions shown in Fig. 1 should be able to provide unambiguously optimized synapomorphies (at least by Fitch optimization in a parsimony analysis that does not integrate a step matrix; Sankoff and Rousseau, 1975) for the clades in question.

Counts for each of the scored-character and character-state distributions for the 189 clades in which the Fitch-parsimony-optimized branches have a minimum length of zero but the SH-like aLRT is ⩾ 0% are presented in Table 9. Our use of A–F in Fig. 1 is identical to our use of A–F in Table 9. The 189 clades may be individually identified by examining the spreadsheet posted as supplemental online data at http://rydberg.biology.colostate.edu/Research/ and cross-referencing the row numbers for the identified clades with the clade numbers assigned on the posted TreeGraph files.

The most common identified scored-character or character-state distribution identified was scenario B (109 clades; 58%) wherein there are shared scored characters between the sister group and only one of the two ingroup terminals (or sub-clades; Table 9). This was followed by scenario E (53–66 clades; 28–35%) wherein there are shared scored characters between the sister group and both ingroup terminals (or sub-clades) but the only potential ingroup synapomorphies are ambiguous based on Fitch optimization. The range listed here for scenario E is because it was difficult to identify the applicable character-state distribution when the focal clade or its sister group contained numerous terminals (Table 9).

Each of the remaining four scenarios accounted for a small minority (⩽6%) of the 189 clades in question (Table 9). Note that at least four of the clades assigned to scenarios B or D may be a consequence of the next clade down in the tree (i.e., the immediately preceding hypothetical ancestor of the clade in question) being contradicted in the SH-like aLRT tree (see Section 3.2 above); the original RAxML tree with bootstrap values was used as there reference topology for these examinations. Still other cases may be a consequence of the next clade down being assigned 0% SH-like aLRT support.

3.6. RAxML branch lengths and branch support

Branch lengths and bootstrap values on the optimal RAxML trees found by the original authors of the 10 sampled supermatrices (with the qualifications noted in Section 2.2 above) are compared to parsimony minimum branch lengths and SH-like aLRT values in Fig. 2. These comparisons were performed to test if those clades that would be considered unsupported by the alternative two criteria are also those with the shortest branch lengths and/or lowest bootstrap values reported by RAxML.

In addition to the pairwise comparisons described above, we also checked whether the apparently unsupported branches (based on parsimony-optimized minimum branch length = 0 or SH-like aLRT = 0) have overlapping 95% confidence intervals for their branch lengths or bootstrap values with the equivalent number of clades that are assigned the shortest RAxML branch lengths while also having a parsimony minimum branch length >0 or that are assigned the shortest RAxML branch lengths while having an SH-like aLRT ≠ 0, respectively. For example, there are 61 clades for the Hinchliff RAxML tree that have a parsimony minimum branch length of zero. The average and ±95% confidence interval for the RAxML branch lengths of these 61 clades was then compared to the average and ±95% confidence interval for the RAxML branch lengths of the 61 clades from the Hinchliff RAxML tree that are the shortest internal branches reported by RAxML while still having a parsimony minimum branch length >0.

Non-overlapping ±95% confidence intervals between the apparently unsupported branches and the supported branches (as identified using parsimony-optimized minimum branch lengths) were identified for at least seven of the 10 sampled supermatrices for both RAxML bootstrap values (Fig. 2A) and RAxML branch lengths (Fig. 2C). Similarly, non-overlapping ±95% confidence intervals between the apparently unsupported branches and the supported branches (as identified using SH-like aLRT values) were identified for all 10 sampled supermatrices for both RAxML bootstrap values (Fig. 2B) and RAxML branch lengths (Fig. 2D). But the results are generally not as clear-cut and favorable to RAxML when comparing the averages and ±95% confidence intervals between the apparently unsupported branches and the equivalent number of apparently supported branches that have the lowest bootstrap values (Fig. 2A and B) or lowest branch lengths (Fig. 2C and D). In these pairwise comparisons the apparently unsupported branches (as identified using parsimony-optimized minimum branch lengths) have higher average RAxML bootstrap values than did the equivalent number of apparently supported branches that have the lowest bootstrap values in six of the 10 supermatrices (only three being significantly higher based on non-overlapping ±95% confidence intervals; Fig. 2A). Furthermore, the branch lengths for these apparently unsupported branches are significantly longer (based on non-overlapping ±95% confidence intervals) than the equivalent number of apparently supported branches that have the shortest branch lengths in nine of the 10 supermatrices (Fig. 2C).

With respect to applying the SH-like aLRT, the apparently unsupported branches were assigned lower bootstrap values for eight of the 10 supermatrices (with non-overlapping ±95% confidence intervals for three supermatrices; Fig. 2B). But the

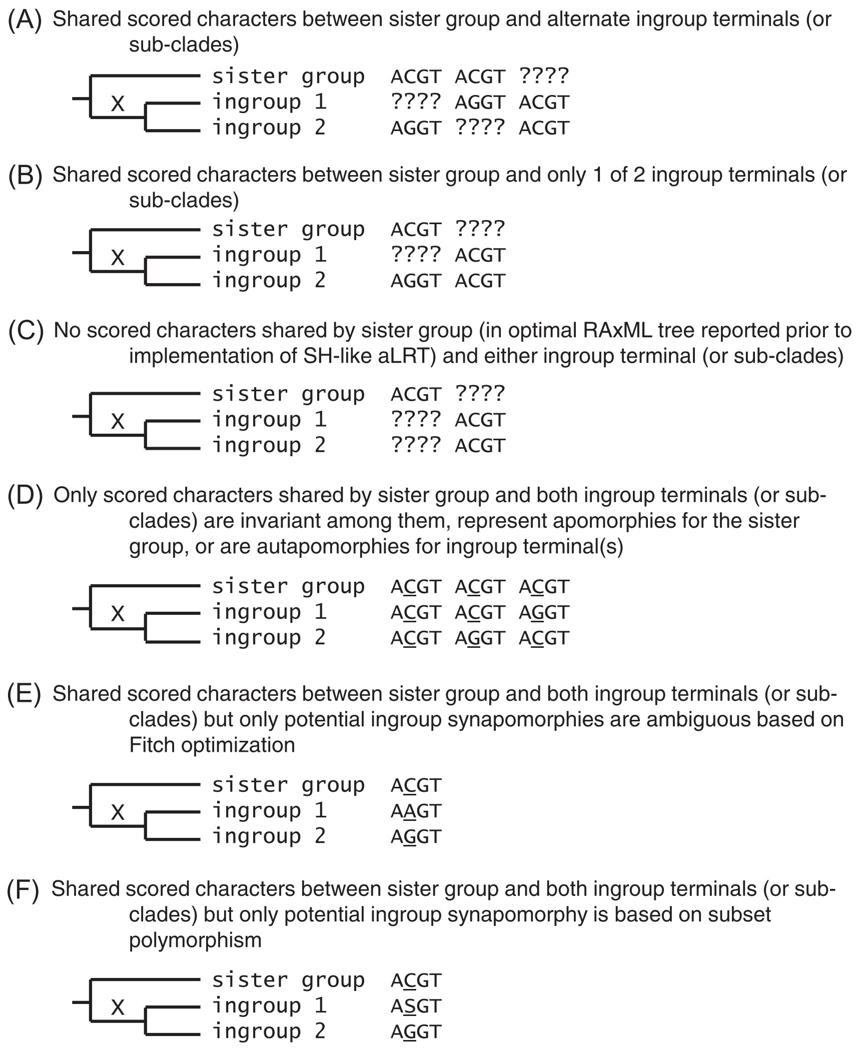
(A) Shared scored characters between sister group and alternate ingroup terminals (or sub-clades)
X
sister group ACGT ACGT ????
ingroup 1 ???? AGGT ACGT
ingroup 2 AGGT ???? ACGT
(B) Shared scored characters between sister group and only 1 of 2 ingroup terminals (or sub-clades)
X
sister group ACGT ????
ingroup 1 ???? ACGT
ingroup 2 AGGT ACGT
(C) No scored characters shared by sister group (in optimal RAxML tree reported prior to implementation of SH-like aLRT) and either ingroup terminal (or sub-clades)
X
sister group ACGT ????
ingroup 1 ???? ACGT
ingroup 2 ???? ACGT
(D) Only scored characters shared by sister group and both ingroup terminals (or sub-clades) are invariant among them, represent apomorphies for the sister group, or are autapomorphies for ingroup terminal(s)
X
sister group ACGT ACGT ACGT
ingroup 1 ACGT ACGT AGGT
ingroup 2 ACGT AGGT ACGT
(E) Shared scored characters between sister group and both ingroup terminals (or sub-clades) but only potential ingroup synapomorphies are ambiguous based on Fitch optimization
X
sister group ACGT
ingroup 1 AAGT
ingroup 2 AGGT
(F) Shared scored characters between sister group and both ingroup terminals (or sub-clades) but only potential ingroup synapomorphy is based on subset polymorphism
X
sister group ACGT
ingroup 1 ASGT
ingroup 2 AGGT

Fig. 1. Scored-character and character-state distributions for cases in which the Fitch-parsimony-optimized branches have a minimum length of zero but the SH-like aLRT is ⩾50%. The sister group, ingroup 1, and/or ingroup 2 terminals shown may each be expanded into clades. The questionably supported ingroup clade is denoted by an X in each case. Hypothetical character-state distributions that are consistent with the patterns observed in the empirical data are shown. Letters A–F match the column headings in Table 9.

Table 9
Counts for each of the scored-character and character-state distributions for cases in which the Fitch-parsimony-optimized branches have a minimum length of zero but the SH-like aLRT is ⩾50%. See Fig. 1 for codes A–F.

Matrix	Lack of scored-character overlap			Character-state distributions			Unknown[a]
	A	B	C	D	E	F	
Bininda	0	2	0	0	0	0	0
Fabre	0	5	1	0	1	0	0
Hedtke	1	12	0	5	11	1	1
Hinchliff	1	14	0	3	7	0	1
Linde	1	4	0	1	0	0	0
Peters	0	0	0	0	10	0	0
Pyron	0	4	0	0	0	0	0
Soltis	1	24	0	1	14	0	11
Springer	1	11	0	0	0	0	0
Wiens	1	28	0	1	10	0	0

[a] Probably code E but difficult to tell because of the numerous terminals in the sister group and/or the ingroup sub-clades.

apparently unsupported branches (as identified using SH-like aLRT values) were generally longer, on average, than the equivalent number of apparently supported branches that have the lowest branch lengths for seven of the 10 supermatrices (only three being significantly longer based on non-overlapping ±95% confidence intervals; Fig. 2D).

3.7. Negative log likelihoods of most parsimonious trees

Negative log likelihoods as calculated by RAxML ver. 7.7.2 for the optimal RAxML tree reported by the original authors (or, in the case of Hinchliff, calculated for this study because the original authors only presented bootstrap majority-rule consensus trees)

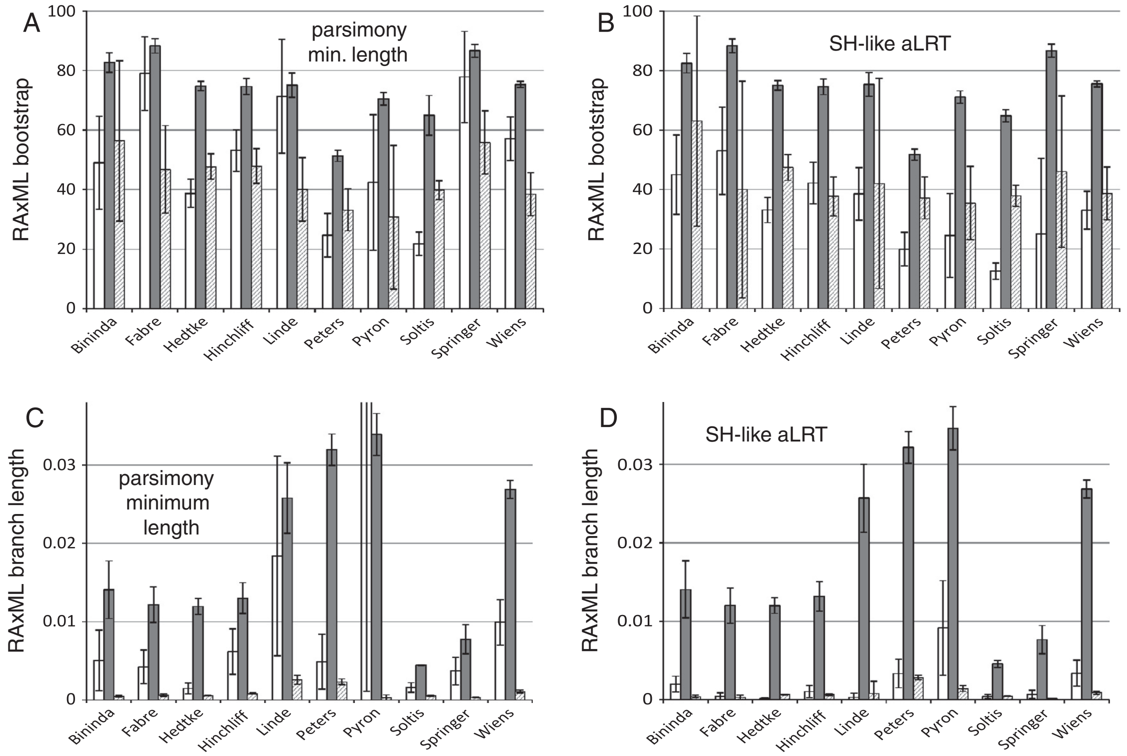

Fig. 2. Comparisons between average RAxML bootstrap values (A and B) and branch lengths (C and D) for each of the ten supermatrices with parsimony minimum branch lengths (A and C) and SH-like aLRT values (B and D). ±95% Confidence intervals are shown. Unshaded bars in A and C are for a parsimony optimized minimum branch length = zero; shaded bars are for a parsimony minimum branch length >0; diagonally striped bars are for those internal branches that are assigned the shortest RAxML branch lengths while also having a parsimony minimum branch length >0. The extreme average branch length and wide ±95% confidence interval for Pyron in C is caused by inclusion of an extreme outlier with an estimated length of 0.237. Unshaded bars in B and D are for SH-like aLRT = 0; shaded bars are for SH-like aLRT ≠ 0; diagonally striped bars are for those internal branches that are assigned the shortest RAxML branch lengths while having an SH-like aLRT ≠ 0.

and up to 10 randomly sampled fully dichotomous most parsimonious trees identified by TNT are presented in Table S2. In all cases except for the Hinchliff supermatrix the parsimony trees are significantly worse at the 1% level based on the SH test. But for the Hinchliff supermatrix, none of the parsimony trees are significantly worse—even at the 5% level. Three of the 10 randomly selected parsimony trees for the Hinchliff matrix have lower negative log likelihoods than that reported by RAxML (based on 100 optimal-tree searches using the GTRCAT model; see Section 2.2 above), but they are not significantly lower.

4. Discussion

4.1. Optimal trees

We infer that our two-stage TNT tree searches probably found most parsimonious trees for the Bininda, Fabre, Hedtke, Hinchliff, Linde, and Springer matrices given that all 10 tree-hybridization analyses for each of these matrices identified shortest known trees of the same length as each other as well as the most parsimonious trees found during the tree-building analyses (Table 3). Yet our TNT tree searches may not have found any most parsimonious trees for the Peters, Pyron, Soltis, and Wiens matrices—particularly for Peters and Soltis given that shorter trees were found during tree hybridization, and even then for only one of the 10 analyses. Ideally the most parsimonious known tree length will be identified numerous times in independent analyses (Davis et al., 2005; Goloboff et al., 2009), but even then there is the concern of being trapped in sub-optimal terraces when analyzing sparse supermatrices (Sanderson et al., 2011).

Our use of TBR-collapsing appears to have been effective in calculating the strict consensus of the most parsimonious known trees for each supermatrix; the strict consensus contains 21–947 fewer clades (average 223) than fully bifurcating trees for the 10 supermatrices (Table 4). Of particular interest is the completely unresolved bush obtained for Soltis despite finding trees of minimum known length in just a single tree-hybridization analysis.

Our compartmentalized analyses and use of the “chkmoves” command in TNT revealed that the properly unsupported resolution obtained by RAxML and the superficial parsimony analyses for Soltis cannot be entirely attributed to a single wildcard terminal that was accidentally left in the published sparse supermatrix. Instead the problem is widespread and occurred despite Soltis’ sampling of 48,465 characters and taking the precautions of only sampling gene alignments scored for ⩾10 terminals, removing spuriously resolved terminals from each gene-alignment tree, and ensuring overlap in terminal sampling between every gene alignment and at least one other gene alignment (Simmons, 2014).

A more effective approach to eliminating problematic terminals in sparse supermatrices than those applied by Soltis is to ensure that the final matrix is decisive for all trees (Sanderson et al., 2010) by ensuring that all terminals are sampled for one or more loci, as done by Hinchliff, Peters, and the Legume Phylogeny Working Group (2013). But even that approach coupled with 16,000+ characters does not ensure that a single fully resolved optimal tree may be justified given that the Hinchliff and Peters strict consensus trees contained 184 and 98, respectively, fewer clades than a fully resolved tree (Table 4).

Still more effective methods of identifying wildcard terminals and properly unsupported resolution are clearly necessary when assembling sparse supermatrices. One approach is to implement Simmons’ (2012b, p. 220) suggested use of group-membership variables (Farris, 1973) wherein a group-membership variable is scored for each clade in the optimal tree for the empirical data.

The distribution of missing data in each set of group-membership variables reflects that of each character partition in the original matrix (e.g., 50 partitions = 50 sets of group-membership variables). A parsimony search is then performed on all group-membership variables (it will run very quickly because there is no character conflict) and the strict consensus is accurately calculated. Any clades in the optimal tree for the empirical data are collapsed if they are not also present in the strict consensus for the group-membership variables.

A second approach is to perform a thorough search for the most parsimonious trees and checking the resolution on the strict consensus (obtained via TBR-collapsing) rather than relying entirely upon RAxML (as was apparently done for nine of the 10 supermatrices sampled here), wherein just a single fully resolved optimal tree is reported. Note that the collapse in resolution reported for the Soltis matrix is not unique to RAxML analyses of sparse supermatrices. In a similar manner, Goloboff (2007) reported that the strict consensus of most parsimonious known trees for the “dense” supermatrix presented by McMahon and Sanderson (2006) is completely unresolved, which they did not discover despite their use of the ratchet (Nixon, 1999) for tree searches.

4.2. Branch-support values

Our finding that summed scaled support was RAxML SH-like aLRT > RAxML bootstrap > superficial parsimony, and in eight of the 10 matrices superficial parsimony > relatively thorough parsimony (Table 4) is consistent with earlier studies that examined two or more of these methods in the context of parsimony (Freudenstein et al., 2004; Müller, 2005; Freudenstein and Davis, 2010; Simmons and Freudenstein, 2011) and/or likelihood (e.g., Pyron et al., 2011; Simmons, 2012a,b, 2014; Simmons and Norton, 2013, 2014). These published comparisons variously represent contrived, simulated, and empirical data.

The most extreme example observed in our study is that for Soltis, wherein the summed scaled support for the SH-like aLRT is 452, that for RAxML bootstrap is 292, that for superficial parsimony bootstrap is 128, and that for the relatively thorough parsimony bootstrap is zero because no clades are present in the strict consensus. Based on the publications cited in the paragraph above as well as our generally consistent results across 10 published sparse supermatrices that represent a broad range of plant and animal taxa, numbers of characters and terminals sampled, and groups of systematists who assembled the matrices, it is clear that one should not rely entirely upon RAxML SH-like aLRT, RAxML bootstrap, or superficial parsimony bootstrap methods to rigorously quantify branch support for sparse supermatrices that contain hundreds or thousands of terminals. Instead, more rigorous and conservative methods are required such as relatively thorough parsimony bootstrap, jackknife, GC, or Bremer-support (Goodman et al., 1985; Bremer, 1988) analyses. Even posterior probabilities generated by Bayesian MCMC analyses may be more conservative than SH-like aLRT or bootstrap values generated by RAxML (Simmons, 2012b, 2014; Simmons and Norton, 2013, 2014; Simmons and Randle, 2014).

4.3. Topological congruence between likelihood and parsimony

The RAxML- and parsimony-derived topologies were generally found to be congruent based on the overall success of resolution and the averaged overall success of resolution between the parsimony topologies and the RAxML clades with ⩾50% bootstrap (Table 7). Certainly not all of the clades were congruent between the parsimony analyses and the RAxML reference tree (19–296 contradictory clades), but these conflicting clades were generally weakly supported by parsimony (ranging from 0 to 14 summed scaled support). These results are largely consistent with Rindal and Brower’s (2011) assertion that few if any mutually well supported conflicting clades are observed between likelihood and parsimony trees in most empirical studies.

Based on the topological congruence between likelihood and parsimony identified by Rindal and Brower (2011) and in this study, we reiterate their conclusion that likelihoodists should not automatically dismiss parsimony analyses of their sparse supermatrices on the grounds that the parsimony analyses are more susceptible to long-branch attraction (Felsenstein, 1978b) and less powerful (Penny et al., 1992). Other factors may be still more important. One fundamental factor that favors thorough parsimony analyses of sparse supermatrices with hundreds to thousands of terminals is being able to distinguish between clades that are unequivocally supported by the data from those that are not. As demonstrated in this and several earlier studies (e.g., Simmons, 2012a,b; Simmons and Norton, 2014; Simmons and Randle, 2014), superficial likelihood analyses that quantify branch support using the bootstrap cannot be relied upon to always make this distinction. The most dramatic example shown here is for the Soltis matrix, which is alternately inferred to have a summed scaled support of 292 (by the RAxML bootstrap) or zero (in the parsimony strict consensus).

4.4. SH-like aLRT = 0 vs. parsimony minimum branch length = 0

The two alternative approaches that we applied to test for properly unsupported clades (i.e., SH-like aLRT = 0 and parsimony minimum branch length = 0) on the most likely reported RAxML trees were in agreement for the majority of the clades in question (Table 8). This general agreement across a diversity of empirical sparse supermatrices reinforces the value of the two approaches. But what about cases where the two approaches disagree—which approach should be relied upon?

For likelihood-based tree-searches, the SH-like aLRT has the clear advantage of being based on the same model and optimality criterion that were used to generate the most likely topology in the first place. Furthermore, just because a branch on a given topology has a positive branch length does not mean that it is necessarily supported by the data. For example, consider a rooted three-terminal clade in which there are only two variable characters, neither of which are variable outside of this clade. Character 1 (A → T) supports the clade of (terminal 1, terminal 2) whereas character 2 (T → A) supports the clade of (terminal 1, terminal 3). Even when analyzed using the highly parameterized GTR Q-matrix, both clades are equally likely. But, regardless of which clade is resolved on the RAxML tree, Fitch optimization will recognize it as a positive-length branch. Hence the parsimony approach fails to recognize the clade as unsupported whereas the SH-like aLRT approach may succeed. So, in cases where the SH-like aLRT equals zero but parsimony reports a positive branch length, the SH-like aLRT result should be accepted. But only a minority of the disagreements shows this pattern (Table 8) and the parsimony approach indicated that many more clades are unsupported. What should one do when the parsimony minimum branch length = 0 and the SH-like aLRT is ⩾50? In these cases the nearest-neighbor-interchange swapping of the SH-like aLRT should not be a limitation because parsimony-optimized branch lengths of zero are always detected by nearest-neighbor-interchange swaps. Therefore, the disagreement may be expected to be caused by use of different optimality criteria.

Based on our examination of the scored-character and character-state distribution for these cases of conflict, 58% of these cases could be attributed to lack of scored characters shared between the sister group of the focal clade and either ingroup terminal (or sub-clades; Table 9). Only a minority of these cases (28–35%) could be attributed to ambiguous Fitch optimization that might be unambiguous (or at least less ambiguous) in a likelihood context. Based on these results we infer that clades for which the parsimony minimum branch length = 0 should be regarded as dubious, even by those who regard likelihood-based optimizations as superior to parsimony.

4.5. RAxML branch lengths and branch support

Based on our results presented in Fig. 2, there is, on average, generally a clear distinction between the apparently unsupported and the supported clades with respect to both RAxML bootstrap values and branch lengths irrespective of which of the two methods is applied to identify apparently unsupported branches. But clades with the shortest branch lengths or lowest RAxML bootstrap values are not necessarily the same as the clades that are apparently unsupported (based on the SH-like aLRT and/or parsimony minimum branch length = 0). Hence there remains the need to identify properly unsupported branches in RAxML analyses of sparse supermatrices rather than just disregarding those clades with low bootstrap support or extremely short branch lengths.

4.6. Negative log likelihoods of most parsimonious trees

Our finding that three of the 10 randomly selected fully bifurcating most parsimonious trees for the Hinchliff matrix have a lower negative likelihood than that found by 100 optimal-tree searches using the GTRCAT model in RAxML provides an empirical proof-of-principle example for Sanderson and Kim's (2000) and Xia's (2006) point that it is possible for thorough parsimony tree searches to find more likely trees than superficial likelihood tree searches. But given that this result was obtained for only one of the 10 sparse supermatrices, and even then the differences in likelihood are not significantly different, we do not expect this to be a general phenomenon—at least when highly parameterized (e.g., GTR + Γ) models are applied.

4.7. Topological incongruence caused by superficial tree searches

Topological incongruence between different optimality criteria and sets of characters can be caused by a range of factors including alignment ambiguity (Morrison and Ellis, 1997), long-branch attraction, nucleotide-frequency heterogeneity among terminals (Lockhart et al., 1992), partial concerted evolution (Sanderson and Doyle, 1992), lineage sorting, introgression, and unrecognized paralogy (e.g., Doyle, 1992). These factors have been thoroughly reviewed and are widely appreciated by systematists. But another factor to consider is superficial tree searches—particularly when applied to ambiguous data (Simmons and Goloboff, 2013; Simmons and Norton, 2014; Simmons and Randle, 2014). Two-stage phylogenetic methods, such as supertrees (Sanderson et al., 1998) and some coalescent analyses (Ané et al., 2007; Degnan et al., 2009; Liu et al., 2010), that use the first-stage trees as input for the second stage may be partially compromised when the first-stage trees contain inaccurate and/or unsupported clades (Gatesy et al., 2002; Townsend et al., 2011; Gatesy and Springer, 2013). In such cases topological incongruence is liable to be overestimated given that there are far more ways for any given group on one tree to be contradicted on another tree than there are ways for the group to be supported on both trees (e.g., Huelsenbeck et al., 2002).

5. Conclusions

Sparse supermatrices that are based mostly or entirely upon publicly available data, such as the 10 sampled here, make an important contribution to phylogenetics by synthesizing existing data across large taxonomic groups with comparatively high taxon coverage relative to the individual component studies for which the initial data were generated. But, given their numerous heuristic shortcuts and less thorough manual curation, we hypothesize that sparse supermatrices are liable to be less accurate *on a per-clade basis* than many of their individual component analyses that include fewer characters and terminals. Therefore we recommend that sparse-supermatrix results be taken into account when making phylogenetic inferences, but not necessarily be regarded as superior to all of their individual component studies when rigor has been sacrificed for speed.

Acknowledgments

We thank two anonymous reviewers for suggestions on how to improve the manuscript; Olaf Bininda-Emonds, Pierre-Henri Fabre, Shannon Hedtke, Cody Hinchliff, Mark Mort, Ralph Peters, Alexander Pyron, Doug Soltis, Mark Springer, and Scott Steppan for providing information and/or sending files from their original phylogenetic analyses; John Gatesy and Andrew Norton for helpful discussions; and Alexandros Stamatakis for a clarification on RAxML.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.06.002.

References

Ané, C., Larget, B., Baum, D.A., Smith, S.D., Rokas, A., 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24, 412–426.

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.

Bininda-Emonds, O.R.P., 2011. Inferring the tree of life: chopping a phylogenomic problem down to size? BMC Biol. 9, 59.

Bininda-Emonds, O.R.P., Stamatakis, A., 2007. Taxon sampling versus computational complexity and their impact on obtaining the tree of life. In: Hodkinson, T.R., Parnell, J.A.N. (Eds.), Reconstructing The Tree of life: Taxonomy and Systematics of Species Rich Taxa. CRC Press, Boca Raton, pp. 77–95.

Bremer, K., 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42, 795–803.

Davis, J.I., Nixon, K.C., Little, D.P., 2005. The limits of conventional cladistic analysis. In: Albert, V.A. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 119–147.

Degnan, J.H., DeGiorgio, M., Bryant, D., Rosenberg, N.A., 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58, 35–54.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., Minh, B.Q., von Haeseler, A., Ebersberger, I., Pass, G., Misof, B., 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily wingless insects. Mol. Biol. Evol. 31, 239–249.

Doyle, J.J., 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. Syst. Bot. 17, 144–163.

Fabre, P.-H., Rodrigues, A., Douzery, E.J.P., 2009. Patterns of macroevolution among primates inferred from a supermatrix of mitochondrial and nuclear DNA. Mol. Phylogenet. Evol. 53, 808–825.

Farris, J.S., 1973. On comparing the shapes of taxonomic trees. Syst. Zool. 22, 50–54.

Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22, 240–249.

Felsenstein, J., 1978a. The number of evolutionary trees. Syst. Zool. 27, 27–33.

Felsenstein, J., 1978b. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Fitch, W.M., 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. 20, 406–416.

Freudenstein, J.V., Davis, J.I., 2010. Branch support via resampling: an empirical study. Cladistics 26, 643–656.

Freudenstein, J.V., van den Berg, C., Goldman, D.H., Kores, P.J., Molvray, M., Chase, M.W., 2004. An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. Am. J. Bot. 91, 149–157.

Gatesy, J., Springer, M.S., 2013. Concatenation versus coalescence versus "concatalescence". Proc. Natl. Acad. Sci. USA 110, E1179.

Gatesy, J., Matthee, C., DeSalle, R., Hayashi, C., 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51, 652–664.

Goloboff, P.A., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

Goloboff, P.A., 2007. Tratamiento de la ambigüedad en grandes matrices de datos. Darwiniana 45, S10–S11.

Goloboff, P.A., Farris, J.S., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.

Goloboff, P.A., Pol, D., 2007. On divide-and-conquer strategies for parsimony analysis of large data sets: Rec-I-DCM3 versus TNT. Syst. Biol. 56, 485–495.

Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., Szumik, C.A., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Goloboff, P.A., Catalano, S.A., Mirande, J.M., Szumik, C.A., Arias, J.S., Källersjö, M., Farris, J.S., 2009. Phylogenetic analysis of 73,060 taxa corroborates major eukaryotic groups. Cladistics 25, 211–320.

Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J., 1985. New perspectives in the molecular biological analysis of mammalian phylogeny. Acta Zool. Fenn. 169, 19–35.

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

Hedtke, S.M., Patiny, S., Danforth, B.N., 2013. The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. BMC Evol. Biol. 13, 138.

Hinchliff, C.E., Roalson, E.H., 2013. Using supermatrices for phylogenetic inquiry: an example using the sedges. Syst. Biol. 62, 205–219.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst. Biol. 51, 673–688.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Murto, H.N. (Ed.), Mammalian Protein Metabolism, vol. 3. Academic Press, New York, pp. 21–132.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649.

Legume Phylogeny Working Group, 2013. Legume phylogeny and classification in the 21st century: progress, prospects and lessons for other species-rich clades. Taxon 62, 217–248.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145.

Liu, L., Yu, L., Edwards, S.V., 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evol. Biol. 10, 302.

Lockhart, P.J., Howe, C.J., Bryant, D.A., Beanland, T.J., Larkum, A.W.D., 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34, 153–162.

Maddison, D.R., 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40, 315–328.

Maddison, D.R., Maddison, W.P., 2001. MacClade: Analysis of Phylogeny and Character Evolution, Version 4.03. Sunderland, Sinauer.

Maddison, W.P., Maddison, D.R., 2013. Mesquite: A Modular System for Evolutionary Analysis. Published by the authors, <http://mesquiteproject.org/mesquite/mesquite.html>.

Maddison, W.P., Donoghue, M.J., Maddison, D.R., 1984. Outgroup analysis and parsimony. Syst. Zool. 33, 83–103.

Malia, M.J., Lipscomb, D.L., Allard, M.W., 2003. The misleading effects of composite taxa in supermatrices. Mol. Phylogenet. Evol. 27, 522–527.

Margush, T., McMorris, F.R., 1981. Consensus *n*-trees. B. Math. Biol. 43, 239–244.

Marshall, D.C., 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. 59, 108–117.

McMahon, M.M., Sanderson, M.J., 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55, 818–836.

Morrison, D.A., 2007. Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. Syst. Biol. 56, 988–1010.

Morrison, D.A., 2013. Evolutionary genomics: statistical and computational methods. Volumes 1 and 2. Syst. Biol. 62, 348–350.

Morrison, D.A., Ellis, J.T., 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. Mol. Biol. Evol. 14, 428–441.

Müller, K., 2005. The efficiency of different search strategies in estimating parsimony jackknife, bootstrap, and Bremer support. BMC Evol. Biol. 5, 58.

Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Nixon, K.C., Carpenter, J.M., 1996. On consensus, collapsibility, and clade concordance. Cladistics 12, 305–321.

Nixon, K.C., Wheeler, Q.D., 1991. Extinction and the origin of species. In: Wheeler, Q.D., Novacek, M. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 119–143.

Nyakatura, K., Bininda-Emonds, O.R.P., 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. BMC Biol. 10, 12.

Penny, D., Hendy, M.D., Steel, M.A., 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7, 73–79.

Peters, R.S., Meyer, B., Krogmann, L., Borner, J., Meusemann, K., Schutte, K., Niehuis, O., Misof, B., 2011. The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. 9, 55.

Pyron, R.A., Wiens, J.J., 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol. Phylogenet. Evol. 61, 543–583.

Pyron, R.A., Burbrink, F.T., Colli, G.R., Montes de Oca, A.N., Vitt, L.J., Kurynski, C.A., Wiens, J.J., 2011. The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees. Mol. Phylogenet. Evol. 58, 329–342.

Rindal, E., Brower, A.V.Z., 2011. Do model-based phylogenetic analyses perform better than parsimony? A test with empirical data. Cladistics 27, 331–334.

Roshan, U.W., Warnow, T., Moret, B.M.E., Williams, T.L., 2004. Rec-I-DCM3: a fast algorithmic technique for reconstructing large phylogenetic trees. Proc. 2004 IEEE Comput. Syst. Bioinform. Conf. 2004, 98–109.

Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.

Sanderson, M.J., 2003. R8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19, 301–302.

Sanderson, M.J., Doyle, J.J., 1992. Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. Syst. Biol. 41, 4–17.

Sanderson, M.J., Kim, J., 2000. Parametric phylogenetics? Syst. Biol. 49, 817–829.

Sanderson, M.J., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109.

Sanderson, M.J., McMahon, M.J., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits of inference. BMC Evol. Biol. 10, 155.

Sanderson, M.J., McMahon, M.M., Steel, M., 2011. Terraces in phylogenetic tree space. Science 333, 448–450.

Sankoff, D., Rousseau, P., 1975. Locating the vertices of a steiner tree in an arbitrary metric space. Math. Program. 9, 240–246.

Schuh, R.T., Polhemus, J.T., 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera). Syst. Zool. 29, 1–26.

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

Siddall, M.E., 2010. Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics 26, 444–452.

Simmons, M.P., 2012a. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol. Phylogenet. Evol. 62, 472–484.

Simmons, M.P., 2012b. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28, 208–222.

Simmons, M.P., 2014. Limitations of locally sampled characters in phylogenetic analyses of sparse supermatrices. Mol. Phylogenet. Evol. 74, 1–14.

Simmons, M.P., Freudenstein, J.V., 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. Mol. Phylogenet. Evol. 61, 177–191.

Simmons, M.P., Goloboff, P.A., 2013. An artifact caused by undersampling optimal trees in supermatrix analyses of locally sampled characters. Mol. Phylogenet. Evol. 69, 265–275.

Simmons, M.P., Norton, A.P., 2013. Quantification and relative severity of inflated branch-support values generated by alternative methods: an empirical example. Mol. Phylogenet. Evol. 67, 277–296.

Simmons, M.P., Norton, A.P., 2014. Divergent maximum-likelihood-branch-support values for polytomies. Mol. Phylogenet. Evol. 73, 87–96.

Simmons, M.P., Randle, C.P., 2014. Disparate parametric branch-support values from ambiguous characters. Mol. Phylogenet. Evol. 78, 66–86.

Simmons, M.P., Webb, C.T., 2006. Quantification of the success of phylogenetic inference in simulations. Cladistics 22, 249–255.

Soltis, D.E., Gitzendanner, M.A., Soltis, P.S., 2007. A 567-taxon data set for angiosperms: the challenges posed by Bayesian analyses of large data sets. Int. J. Plant Sci. 168, 137–157.

Soltis, D.E., Mort, M.E., Latvis, M., Mavrodiev, E.V., O'Meara, B.C., Soltis, P.S., Burleigh, J.G., Rubio de Casas, R., 2013. Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. Am. J. Bot. 100, 916–929.

Springer, M.S., Meredith, R.W., Gatesy, J., Emerling, C.A., Park, J., Rabosky, D.L., Stadler, T., Steiner, C., Ryder, O.A., Janecka, J.E., Fisher, C.A., Murphy, W.J., 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS ONE 7, e49521.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

Stamatakis, A., 2014. The RAxML v8.0.X manual. <http://sco.h-its.org/exelixis/resource/download/NewManual.pdf>, (downloaded 04.03.14).

Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–771.

Stöver, B.C., Müller, K.F., 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11, 7.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a python library for phylogenetic computing. Bioinformatics 26, 1569–1571.

Sundberg, K., Carroll, H., Snell, Q., Clement, M., 2008. Incomparability of results between phylogenetic search programs. In: Proceedings of the 2008 International

Conference on Bioinformatics and Computational Biology (BIOCOMP'08), pp. 81–84.

Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis Using Parsimony (*And Other Methods). Sinauer Associates, Sunderland.

Townsend, T.M., Mulcahy, D.G., Noonan, B.P., Sites, J.W., Kuczynski, C.A., Wiens, J.J., Reeder, T.W., 2011. Phylogeny of iguanian lizards inferred from 29 nuclear loci, and a comparison of concatenated and species-tree approaches for an ancient, rapid radiation. Mol. Phylogenet. Evol. 61, 363–380.

van der Linde, K., Houle, D., Spicer, G.S., Steppan, S.J., 2010. A supermatrix-based molecular phylogeny of the family Drosophilidae. Genet. Res. 92, 25–38.

Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60, 719–731.

Xia, X., 2006. Molecular phylogenetics: mathematical framework and unsolved problems. In: Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M. (Eds.), Structural Approaches to Sequence Evolution. Springer, New York City, pp. 171–191.

Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol. Biol. Evol. 14, 717–724.

Zwickl, D.J., 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion. Ph.D. Dissertation, The University of Texas at Austin.