The effects of subsampling gene trees on coalescent methods applied to ancient divergences

Mark P. Simmons, Daniel B. Sloan, John Gatesy

Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
Department of Biology, University of California, Riverside, CA 92521, USA

ARTICLE INFO

Article history:
Received 9 October 2015
Revised 3 December 2015
Accepted 20 December 2015
Available online 6 January 2016

Keywords:
ASTRAL
Concatenation
MP-EST
Phylogeny
Shortcut coalescent methods
STAR

ABSTRACT

Gene-tree-estimation error is a major concern for coalescent methods of phylogenetic inference. We sampled eight empirical studies of ancient lineages with diverse numbers of taxa and genes for which the original authors applied one or more coalescent methods. We found that the average pairwise congruence among gene trees varied greatly both between studies and also often within a study. We recommend that presenting plots of pairwise congruence among gene trees in a dataset be treated as a standard practice for empirical coalescent studies so that readers can readily assess the extent and distribution of incongruence among gene trees. ASTRAL-based coalescent analyses generally outperformed MP-EST and STAR with respect to both internal consistency (congruence between analyses of subsamples of genes with the complete dataset of all genes) and congruence with the concatenation-based topology. We evaluated the approach of subsampling gene trees that are, on average, more congruent with other gene trees as a method to reduce artifacts caused by gene-tree-estimation errors on coalescent analyses. We suggest that this method is well suited to testing whether gene-tree-estimation error is a primary cause of incongruence between concatenation- and coalescent-based results, to reconcile conflicting phylogenetic results based on different coalescent methods, and to identifying genes affected by artifacts that may then be targeted for reciprocal illumination. We provide scripts that automate the process of calculating pairwise gene-tree incongruence and subsampling trees while accounting for differential taxon sampling among genes. Finally, we assert that multiple tree-search replicates should be implemented as a standard practice for empirical coalescent studies that apply MP-EST.

1. Introduction

Coalescent methods that allow for differential lineage sorting among genes are now often applied in phylogenetic analyses of both recently diverged and ancient lineages—even in cases where there is no reason to believe that the anomaly zone, wherein the most likely gene-tree topology contradicts the phylogenetic tree (Degnan and Rosenberg, 2006), may apply. Shortcut coalescent methods (i.e., those that do not co-estimate the phylogenetic tree with the gene trees; Gatesy and Springer, 2014) are statistically consistent if the gene trees are known without error (Liu et al., 2009, 2010; Mirarab et al., 2014). Gene-tree-estimation error can be caused by limited character variation among recently diverged lineages (Huang and Knowles, 2009), but for ancient lineages there are the additional potential problems of long-branch attraction (Felsenstein, 1978), mis-rooting (Rosenfeld et al., 2012), convergent nucleotide composition (Lockhart et al., 1992), and short coalescent genes (Hobolth et al., 2011; Gatesy and Springer, 2014; Springer and Gatesy, 2016). Indeed, high levels of gene-tree conflict are frequently reported in empirical studies that sampled ancient lineages (e.g., Betancur-R et al., 2013; Salichos and Rokas, 2013; Pyron et al., 2014).

Gene-tree incongruence that is caused by estimation errors can be difficult to distinguish from the biological process of lineage sorting (Yang, 2002; Leigh et al., 2008; Betancur-R et al., 2014). This is particularly true when applying coalescent methods to resolve rapid ancient radiations because resolution of such problems is dependent upon rapidly evolving genes to provide synapomorphies along these short branches, yet these same synapomorphies are likely to be obscured by subsequent mutations (Maddison and Knowles, 2006). Mis-estimated gene trees have been identified as a probable cause of artifacts in shortcut
A variety of approaches have been proposed to quantify and/or help minimize phylogenetic-inference artifacts caused by divergent gene trees. Leigh et al. (2011a) clustered gene trees based on their shared bipartitions, after which each cluster may be analyzed independently of the others. De Vienne et al. (2012) subsampled both genes and taxa with the most similar pairwise distances among taxa in their gene trees. Salichos and Rokas (2013) subsampled those gene trees with high bootstrap support (Felsenstein, 1985; Mirarab et al. 2014) excluded two outlier genes with high pairwise Robinson-Foulds distance (hereafter “RF distance”; Robinson and Foulds, 1981) relative to other gene trees (clades in gene trees with <75% bootstrap support were collapsed). Similarly, Pyron et al. (2014) quantified pairwise incongruence among gene trees and also between each gene tree and the coalescent phylogenetic tree using RF distances. The latter approach is implemented in STRAW (Shaw et al., 2013), which also takes into account differential taxon sampling among gene trees. Sharma et al. (2014) alternatively subsampled genes with the lowest percentage of missing terminals or those with the highest percent pairwise amino-acid identity. Zimmermann et al. (2014) used aBEAST (Heled and Drummond, 2010), a coalescent method that co-estimates gene trees and the phylogenetic tree, to improve gene-tree estimation prior to applying a shortcut coalescent method.

Of these various alternatives, we focused on the approach of subsampling those gene trees that have low average RF distance with other gene trees after correcting for the number of shared terminals (hereafter the “RF method”). By comparing gene trees pairwise, the RF method bypasses comparison to a species tree inferred from the gene trees and does not favor similarity to an initial species tree that may be biased by outlier gene trees. This same approach can be implemented using rooted triplets or unrooted quartets (Estabrook et al., 1985; Leigh et al., 2011b; Zwickl et al., 2014) instead of RF, but these methods, despite their expected greater stability to outlier terminals, have not performed well in practice (Kuhner and Yamato, 2015).

In this study, we sampled eight empirical studies of ancient lineages with diverse numbers of taxa and genes for which the original authors applied one or more coalescent methods. For each of these studies we quantified topological incongruence among gene trees, compared the relative performance of three shortcut coalescent methods (ASTRAL, MP-EST, and STAR) that are frequently applied to empirical datasets, tested the effectiveness of subsampling gene trees using the RF method for improving coalescent-based phylogenetic inference, quantified how heuristic MP-EST tree searches can affect the inferred phylogenetic tree, tested alternative character-coding and character-sampling approaches for two studies, and used the RF method to identify outlier gene trees. We did so in a particularly challenging context—ancient lineages for which we expect gene-tree-estimation error to be severe.

### 2. Materials and methods

#### 2.1. Primary empirical studies sampled

For the core analyses of our study, we sampled the following eight empirical studies: Betancur-R et al. (2013; hereafter “Betancur”), Chiari et al. (2012; hereafter “Chiari”), McCormack et al. (2012; hereafter “McCormack”), Pyron et al. (2014; hereafter “Pyron”), Townsend et al. (2011; hereafter “Townsend”), Wiens et al. (2012; hereafter “Wiens”), Xi et al. (2014; hereafter “Xi”) and Zhong et al. (2013; hereafter “Zhong”). These studies include diverse animal and plant lineages, numbers of taxa (16–214), and numbers of gene trees (20–333; Table 1).

Gene trees and concatenation-based phylogenetic trees were obtained from the authors, downloaded from Dryad, or manually re-created from the original authors’ figures. In most cases, the original authors’ gene trees and phylogenetic trees were used, though some gene trees required re-rooting (applicable to Betancur wherein gene trees were rooted using Zeus when the two out-groups [Polymixia and Zeus] were resolved as polyphyletic, and Zhong for which gene trees were rooted using Micromonas when the three Chlorophyte outgroups were not resolved as a clade) and/or re-naming a small minority of inconsistently named taxa. We used Simmons and Gatesy’s (2015) concatenation-based phylogenetic trees for Xi, which were based on partitioned-by-gene RAXML tree searches using 100 search replicates, in contrast to the original authors’ unpartitioned analysis from a single RAxML search replicate. We also used Simmons and Gatesy’s (2015) gene trees that were based on subsamples of the characters because Xi were not able to provide these.

In cases where the original authors analyzed alternative datasets with different numbers of gene trees and/or taxa, we always selected the dataset with the higher number of taxa (applicable to McCormack, Townsend, and Wiens), even if the original authors did not apply coalescent analyses to these datasets (applicable to Townsend and Wiens). For Zhong, wherein the original authors analyzed the same number of taxa with different numbers of genes (184 or 289), we selected the larger dataset. All gene trees and concatenation-based phylogenetic trees used are posted as supplemental online data at: http://dx.doi.org/10.6084/m9.figshare.1615928.

Chiari performed their coalescent (and concatenation) analyses alternatively using gene trees inferred from amino-acid or nucleotide characters (hereafter Chiari AA and Chiari DNA). Both concatenation-based analyses and the Chiari AA coalescent analysis resolved the topology (((turtles)(crocodylids)(birds)),

<table>
<thead>
<tr>
<th>Study</th>
<th># Terminals</th>
<th># Gene trees</th>
<th>Lineage*</th>
<th>Reference clade(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Betancur-R et al. (2013)</td>
<td>214</td>
<td>20</td>
<td>Percomorpha</td>
<td>Flatfishes</td>
</tr>
<tr>
<td>Chiari et al. (2012)</td>
<td>16</td>
<td>248</td>
<td>Sarcopterygi</td>
<td>((turtles)(crocodylids)(birds))</td>
</tr>
<tr>
<td>McCormack et al. (2012)</td>
<td>29</td>
<td>183</td>
<td>Anniota</td>
<td>(Rodentia)(Lagomorpha)</td>
</tr>
<tr>
<td>Pyron et al. (2014)</td>
<td>33</td>
<td>333</td>
<td>Squamata</td>
<td>Caenophidia</td>
</tr>
<tr>
<td>Townsend et al. (2011)</td>
<td>76</td>
<td>29</td>
<td>Lepidosauria</td>
<td>((Leioosaurida)(Oplurids))</td>
</tr>
<tr>
<td>Wiens et al. (2012)</td>
<td>171</td>
<td>44</td>
<td>Anniota</td>
<td>(Dibamidae)(Gekkota)*</td>
</tr>
<tr>
<td>Xi et al. (2014)</td>
<td>46</td>
<td>310</td>
<td>Tracheophyta</td>
<td>(Amborella, (other flowering plants))</td>
</tr>
<tr>
<td>Zhong et al. (2013)</td>
<td>32</td>
<td>289</td>
<td>Viridiplantae</td>
<td>((Zygnematales)(land plants))</td>
</tr>
</tbody>
</table>

* Including outgroup(s).

* Together as a clade or as a paraphyletic group on successive branches.
whereas the Chiari DNA coalescent analysis resolved the contrary topology (((turtles)(crocodylians))(birds)). We sampled both the Chiari AA and Chiari DNA gene trees in our analyses.

Xi performed their coalescent (and concatenation) analyses alternatively using gene trees inferred from all characters or subsets of the characters identified using Observed Variability (OV; Goremynkin et al., 2010) or Tree Independent Generation of Evolutionary Rates (TIGER; Cummings and McInerney, 2011). OV and TIGER are methods that are designed to separate out slowly evolving characters, which may be expected to be more reliable for inferring deep divergences, from faster evolving characters (Goremynkin et al., 2010; Cummings and McInerney, 2011). Xi used OV and TIGER independently of each other to split the parsimony-informative characters into halves (based on all genes together rather than for each individual gene) and then analyzed each half together with all parsimony-uninformative characters after the characters were parsed back to the various genes. The nominally fast evolving characters were labelled OV fast and TIGER fast, whereas the nominally slow evolving characters were labelled OV slow and TIGER slow. Xi favored their OV-slow and TIGER-slow analyses. We sampled all five of these sets of gene trees (all characters included, OV fast, OV slow, TIGER fast, and TIGER slow).

2.2. Reference clades

For each of the eight empirical studies, one or more reference clades were selected to test how well each of the three coalescent methods performed in supporting or contradicting the clade(s) (Table 1). The reference clades generally were selected as well supported natural groups. In most cases, the focal clade(s) are among the primary foci of the original study.

The flatfishes (i.e., Pleuronectiformes) are the reference clade for Betancur. Flatfish monophyly has been subsequently supported by Betancur-R and Ortí (2014; but see Campbell et al., 2014).

The five reference clades for Chiari are: (((turtles)(crocodylians))(birds)). In addition to including the primary focus of the original paper, these clades were selected because the original authors reported that all five of them were supported by their DNA- and AA-based concatenation analyses. Likewise, ASTRAL analyses of both sets of gene trees gave the same result (Mirarab et al., 2014). MP-EST analysis of the AA-based gene trees supported the same five clades, but MP-EST analysis of the DNA-based gene trees supported an alternative topology. The five reference clades were also supported by the phylogenomic data of Crawford et al. (2012).

For McCormack, the obvious potential reference clades are Afrotheria sister to the rest of the placental mammals, but Gatesy and Springer (2014) demonstrated that this result is only ambiguously supported by McCormack’s own data and is strongly contradicted by other studies (e.g., Song et al., 2012). Instead we used the conventional clades (Lagomorpha, (Rodentia)), which were supported by McCormack, Song et al. (2012), and many previous studies (summarized in Gatesy and Springer, 2014).

The clade Caenophidia, which is Pyron’s ingroup, is the reference clade for that study. This clade has been supported in previous studies (Wiens et al., 2008; Pyron et al., 2013). Although supported in four of Pyron’s seven analyses, Caenophidia was contradicted by neighbor joining (Saitou and Nei, 1987)-based concatenation analysis as well as two shortcut coalescence methods, Njst (Liu and Yu, 2011) and STAR. Hence this clade serves as a potentially severe test of the alternative coalescent methods.

The three reference clades for Townsend are: (((Leiosaurids) (Oplurids))). These clades were selected because Townsend noted that the clade (Leiosaurids, Oplurids) was effectively unsupported by many gene trees as well as the BEST-based coalescent analysis (Liu, 2008). Hence this clade is a potentially severe test of the alternative coalescent methods. The three reference clades were also subsequently supported by Blankers et al. (2013), who included improved taxon sampling.

Two of the reference clades for Wiens are Dibamidae and Gekkota. Wiens resolved the clade (Dibamidae, Gekkota) in their coalescent- and coalescent-based inferred phylogenies and cited this as a surprising result in their abstract. Indeed, previous studies had resolved Gekkota and Dibamidae as two successive clades that together form a paraphyletic group with either Gekkota (Townsend et al., 2004; Hugall et al., 2007) or Dibamidae (Vidal and Hedges, 2005; Lee, 2009; Wiens et al., 2010 [for their parsimony-based concatenated analysis but not their Bayesian-based concatenated analysis]) resolved as the first lineage. Hence as a reference we accepted all three of these alternative resolutions of Dibamidae and Gekkota relative to each other. Note that this decision favored ASTRAL over STAR because ASTRAL often resolved the two lineages as successive sisters in a paraphyletic group, whereas STAR only did so rarely in our Wiens analyses.

The two reference clades for Xi are (Amborella, (other flowering plants)). Although the support for Amborella as the sole extant sister group to the rest of the flowering plants may be considered controversial (e.g., Drew et al. (2014), Ruhfel et al. (2014), Wickett et al. (2014) and Simmons and Gatesy (2015) versus Xi et al. (2014) and Goremynkin et al. (2015)), Simmons and Gatesy (2015) demonstrated that this is the better supported resolution for Xi’s data by both concatenation and coalescent analyses.

Finally, the three reference clades for Zhong are ((Zygnermatales)(land plants)). The sister-group relationship between Zygnermatales and the land plants was the focus of Zhong, and has also been corroborated by Zhong et al. (2014) and Wickett et al. (2014). However, Springer and Gatesy (2014) pointed out the limitations of Zhong’s dataset given the correlated patterns of missing data among gene trees, divergent gene-tree topologies that are almost certainly artifacts rather than cases of lineage sorting, and the contrary phylogenetic relationships inferred by MP-EST versus STAR. Hence the sister-group relationship between Zygnermatales and the land plants may be a severe test of the alternative coalescent methods.

In all studies except Betancur, each inferred coalescent species tree was scored as zero if the topology included all of the reference clade(s) and was scored as one if the topology contradicted one or more of the reference clades. For Betancur, wherein the single reference clade is flatishes, a more sensitive measure was applied to better account for the diverse set of topologies observed. The scoring measure is based on Fitch (1971) optimization of a group-membership character (Farris, 1973) onto the inferred topology to quantify the number of inferred lineages that contradict flatfish monophyly. For example, a score of one indicates that the flatfish are polyphyletic and separated into two separate clades or that a single non-flatfish lineage is nested within the flatfish clade.

2.3. Subsampling gene trees

To calculate the pairwise RF topological-only distance between gene trees, we modified Pankey’s (2014) Python script posted at https://scriptomika.wordpress.com/2014/01/27/59/. Importantly for empirical studies, Pankey’s script accounts for differential taxon sampling among gene trees by deleting all taxa that are unique to either tree in each pairwise comparison. We refer to the modified script, which is dependent upon p4-phylogenetics (https://github.com/pgp/p4-phylogenetics) and DendroPy (Sukumaran and Holder, 2010), as “RF Distances Filter.” RF Distances Filter is posted as supplemental online data at: http://dx.doi.org/10.6084/m9.figshare.1615928.

RF Distances Filter calculates and reports all pairwise RF distances for a set of gene trees as well as the number of shared terminals in each gene-tree comparison. It then calculates and reports
the average RF distance for each input gene tree relative to all other input gene trees after using the scaling function (RF distance)/
(2 × ((number of shared terminals) – 3)), as described by
Rosenberg and Kumar (2001). Note that this scaling function can result in inflated estimates of congruence for trees that contain few (<10) terminals simply because there are fewer possible topologies such that the topologies are more likely to be congruent by chance alone (Zwickl and Hillis, 2002). Zwickl and Hillis (2002)
noted that the scaling function may be corrected for these instances by adjusting the RF distance based on the expected RF distance between randomly selected trees.

Finally, RF Distances Filter outputs a set of text files of gene
trees at user-defined percentage intervals. The default is for RF Dis-
tances Filter to do so based on the scaled average RF distances for
each gene tree. For example, 90% subsampling would remove the
10% of gene trees that are, on average, least congruent in pairwise
comparisons to all other gene trees in the original dataset. Alterna-
tively, RF Distances Filter can also output sets of gene trees based
on random orderings of all gene trees (i.e., without respect to pair-
wise incongruence). In both cases each more inclusive subsam-
pling includes all trees in the more restrictive subsampling file
(e.g., all trees in the 10% subsampling file are also included in the
20%, 30%, . . . 90% subsampling files). When applicable, RF Distances
Filter always rounds down. For example, 10% subsampling for the
Townsend et al. (2011) set of 29 gene trees is treated by RF Dis-
tances Filter as two genes rather than three genes. We used RF Dis-
tances Filter to subsample gene trees from each study at nine
subsamples as two genes rather than three genes. We used RF Distances Filter together with Microsoft Excel to calculate the
scaled RF distance between the inferred trees and the reference
topology reported by the original authors (with the exception of
Xi; see Section 2.1 above) as the reference tree. Because all eight of the studies sampled include ancient lineages, we expect that
gene-tree-estimation error (caused by, for example, long-branch
attraction, saturation, weak phylogenetic signal, model mis-
specification, differential base composition among taxa) is a severe
problem for coalescent methods (Bayzid and Warnow, 2013;
Salichos and Rokas, 2013; Springer and Gatesy, 2014)—perhaps
even more so than is lineage sorting for concatenation methods
at these deep branches (Patel et al., 2013; Gatesy and Springer,
2014; Simmons and Gatesy, 2015). Thus, measuring congruence
of the coalescent trees to the concatenation result for the full data-
set is justified.

For the second approach we used the optimal coalescent tree
identified when all gene trees were sampled. This approach was
implemented to be specific to each coalescent method (e.g.,
the 10% subsampling ASTRAL tree was compared to the 100% sampling
ASTRAL tree whereas the 10% subsampling MP-EST tree was com-
pared to the 100% sampling MP-EST tree). Our null expectation is that
the tree based on all available data is the most robust hypoth-
esis with the greatest explanatory power (Kluge, 1989; Nixon and
Carpenter, 1996; but see Leconte and Deleporte, 2005). As such,
comparing the performance of methods when applied to subsets of
the available data relative to the same method when applied to
the full set of data is commonplace (e.g., Cummings et al., 1995;
Simmons and Freudenstein, 2003; Miller and Hormiga, 2004).

Given that we conducted up to 1000 MP-EST tree searches for
each set of gene trees, we evaluated the thoroughness of the
MP-EST searches by quantifying the scaled RF distance between all output species trees from the various searches and the single optimal species-tree topology by using a modified version of RF Distances Filter. We also report the number of times that the MP-EST tree searches found trees with the highest known likelihood.

2.6. Using the RF method to identify outlier gene trees

Song et al. (2012; hereafter “Song”) executed a phylogenomic coalescent analysis of 447 genes sampled from 36 mammalian taxa and a bird outgroup. Springer and Gatesy (2016) criticized Song for basing their coalescent analyses on duplicated genes, genes with transposed taxon names (also see Mirarab et al., 2014), data matrices with obvious alignment errors, and low quality gene-tree searches using a mistaken substitution model. After addressing these concerns, Springer and Gatesy (2016) reduced Song’s original 447 gene trees to a curated set of 413 improved gene trees. As an intermediate step, Springer and Gatesy (2016) also re-analyzed the 439 unique genes in Song using the GTR + I model with SPR + NNI branch swapping in PhyML (Guindon et al., 2010) to compare the results with Song’s original gene trees, which apparently were inferred using the GTR model with NNI branch swapping in PhyML. In the set of 439 genes, the eight duplicated loci in Song’s data set were deleted, but 26 genes that were misaligned and 21 genes with swapped taxon names were not fixed. We compared scaled pairwise RF distances among these three sets of gene trees as a means to quantify Springer and Gatesy’s (2016) improvements. We also used RF Distances Filter to assess the utility of the RF method for pinpointing additional, highly conflicting gene trees that were not identified by eye in Springer and Gatesy (2016).

3. Results

3.1. Average pairwise distances among gene trees

The raw data for all results are posted as an Excel file in supplemental online data at: http://dx.doi.org/10.6084/m9.figshare.1615928. The average scaled pairwise RF distances among gene trees for each of the eight empirical studies is presented in Fig. 1. None of the studies included gene trees with an average scaled pairwise RF distance of <0.3. That is, all gene trees from all eight studies conflict, on average, with >30% of the clades in all other gene trees from the same study. Wiens and Townsend included gene trees with relatively low average scaled pairwise RF distances (averages of 0.41 and 0.43, respectively), whereas McCormack and Zhong had extremely high average scaled pairwise RF distances (averages of 0.85 and 0.73, respectively). Several studies had long right-hand tails to their distributions of average scaled pairwise RF distances, including Chiari (both AA- and DNA-based gene trees), Pyron, and Xi (all characters, OV-fast, and TIGER-fast), which indicate highly incongruent outlier gene-tree topologies.

In comparing the character-coding strategies applied by Chiari, coalescent analyses using randomly subsampled AA-based gene trees consistently outperformed those using randomly subsampled DNA-based gene trees (Fig. 2). But when the RF method was applied, coalescent analyses using both AA- and DNA-based gene trees almost always resolved all five reference clades. In comparing the character-sampling strategies applied by Xi when the RF method was applied (random subsamples were not generated for the OV- or TIGER-based analyses), coalescent analyses of gene trees inferred using all characters consistently outperformed both the AA- and TIGER-based analyses (Figs. 2, S1).

3.2. Reference clades

The fractions of trees that are contrary to the reference clade(s) for each of the eight empirical studies as inferred by the three coalescent methods tested under different percentage subsamplings of the gene trees are presented in Fig. 2. In comparing performance among the different coalescent methods when applied to random subsamples of gene trees, no single coalescent method universally performed the best. But the closest method to do so was ASTRAL, which generally outperformed both MP-EST and STAR for Betancur, Chiari DNA, Xi all characters, and Zhong. ASTRAL generally outperformed STAR for Chiari AA, Pyron, and Wiens. It also generally outperformed MP-EST for McCormack. But ASTRAL did not clearly outperform MP-EST for Chiari AA or Pyron, and did not clearly outperform STAR for McCormack. The only study wherein STAR clearly outperformed both ASTRAL and MP-EST was Townsend. MP-EST did not clearly outperform both ASTRAL and STAR for any study (but note that it was not applied to random subsamples of gene trees for either Betancur or Wiens because of computational limitations).

In comparing performance of the RF method to randomly subsampling gene trees, the RF method generally performed better for one or more coalescent method when applied to Betancur (ASTRAL), Chiari AA (ASTRAL, MP-EST), Chiari DNA (ASTRAL, MP-EST, STAR), McCormack (ASTRAL, STAR), Pyron (STAR), Townsend (ASTRAL, MP-EST, STAR), Wiens (STAR), and Xi all characters (MP-EST, STAR, Fig. 2). Yet the RF method was generally outperformed by random subsamples for STAR when applied to Betancur and Zhong.

Subsampling using the RF method outperformed the 100% sampling results for three or more subsamplings when applied to Betancur (ASTRAL only), Chiari DNA (MP-EST only), McCormack (MP-EST only), Xi all characters (MP-EST only), Xi OV-fast characters (ASTRAL only), and Xi TIGER-fast characters (ASTRAL only; Figs. 2, S1). By contrast, the 100% sampling results consistently outperformed subsampling using the RF method for Betancur (STAR only).

In comparing the character-coding strategies applied by Chiari, coalescent analyses using randomly subsampled AA-based gene trees consistently outperformed those using randomly subsampled DNA-based gene trees (Fig. 2). But when the RF method was applied, coalescent analyses using both AA- and DNA-based gene trees almost always resolved all five reference clades. In comparing the character-sampling strategies applied by Xi when the RF method was applied (random subsamples were not generated for the OV- or TIGER-based analyses), coalescent analyses of gene trees inferred using all characters consistently outperformed the OV-slow and TIGER-slow selected characters (Figs. 2, S1).

3.3. Congruence with concatenation: ASTRAL vs. MP-EST vs. STAR

Average scaled RF distances for ASTRAL, MP-EST, and STAR results (all gene trees as well as nine subsamplings) relative to the concatenation tree are presented in Fig. 3A. These subsamplings are based on application of the RF method. ASTRAL-based coalescent trees had the lowest average scaled RF distances to the concatenation trees for Betancur, Chiari DNA, Wiens, all four Xi character-subsampling approaches, and Zhong. By contrast, MP-EST had lowest average scaled RF distances to the concatenation trees for Chiari AA and Townsend. Only for Pyron did STAR-based coalescent trees have lowest average scaled RF distances, but the differences in average scaled values were marginal, just 0.007 for ASTRAL and 0.003 for MP-EST.

3.4. Internal consistency: ASTRAL vs. MP-EST vs. STAR

ASTRAL performed better still when the average scaled RF distances were compared to the tree inferred by the same coalescent method when 100% of the gene trees were sampled (Fig. 3B), which is a measure of self-consistency. Here ASTRAL-based coalescent
trees had the lowest average scaled RF distances for all studies except Pyron and Townsend. As before, MP-EST performed the best for Townsend (by just 0.006) while STAR performed the best for Pyron (by 0.037).

The relative ranking of ASTRAL vs. MP-EST and STAR were similar when the subsamplings were based on random subsets of gene trees, albeit somewhat less favorable for ASTRAL (Fig. S2). For these comparisons, ASTRAL was outperformed by MP-EST and/or STAR for Chiari AA, McCormack, and Pyron when assessing average scaled RF distances to the concatenation tree, as well as Chiari AA, Chiari DNA, Pyron, and Townsend for averaged scaled RF distances to the tree identified by the same method when 100% of gene trees were sampled. Note that data are not available in Fig. S2 for any of the Xi character-subsampling approaches nor MP-EST for Betancur and Wiens.

3.5. Congruence with concatenation: RF method vs. random subsamplings

To compare the performance of the RF method relative to the average of random subsamplings, scaled RF distances for ASTRAL, MP-EST, and STAR relative to the concatenation tree are presented in Fig. S1. In this context the RF method consistently outperformed random subsamples for Betancur (ASTRAL), McCormack (MP-EST, ASTRAL), Townsend (ASTRAL), Xi all characters (MP-EST), and Zhong (ASTRAL, MP-EST, STAR). By contrast, random subsamples only consistently outperformed the RF method for Betancur (STAR).

Subsampling using the RF method outperformed the 100% sampling results for three or more subsamplings when applied to Betancur (MP-EST), McCormack (MP-EST), Pyron (STAR), Wiens
By contrast, the 100% sampling results consistently outperformed subsampling using the RF method for Betancur (ASTRAL, STAR), Chiari DNA (STAR), and Townsend (STAR).

### 3.6. Internal consistency: RF method vs. random subsamplings

For a final comparison of the performance of the RF method relative to the average of random subsamplings, scaled RF distances for ASTRAL, MP-EST, and STAR relative to the tree inferred by the same coalescent method when 100% of the gene trees were sampled are presented in Fig. 54. In this context the RF method only consistently outperformed random subsamples for Betancur (ASTRAL, STAR), Chiari DNA (STAR), and Townsend (STAR).

### 3.7. MP-EST tree searches

The number of MP-EST searches for which the highest pseudo-likelihood score was found for each of the empirical studies is presented in Table S1. Of the 1000 MP-EST searches that were performed for each of the 100 McCormack, 100 Pyron, 100 Xi all characters, and 100 Zhong datasets, the optimal known tree was identified more than once (up to four times) for 20 datasets (2 McCormack, 4 Pyron, 1 Xi all characters, 13 Zhong). Similarly, among the average of 928 MP-EST searches that were performed for each of the 100 Townsend datasets, optimal known trees were identified more than once (up to three times) for four of the datasets. By contrast, the optimal known tree was identified 2–914 times for 184 of the 200 Chiari AA/DNA matrices, which include just 16 taxa. The general pattern observed for the Chiari matrices is that optimal known trees were identified more frequently when fewer gene trees were sampled (data not shown).
Scaled RF distances between the optimal tree found by MP-EST and trees from all MP-EST searches for each of the 10 percentages of gene trees sampled are presented in Fig. 4 (with subsamplings of gene trees selected using the RF method) and Fig. S5 (with subsamplings of gene trees selected using a set of random subsamples). In both cases the scaled RF distances are averaged over blocks of 50 trees.

Of the 130 sets of searches for which subsamplings of gene trees were selected using the RF method, the optimal tree topology was identified in all searches for 28 sets, 95–99% of searches for eight sets, and 90–94% of searches for six sets. At the other extreme, the optimal known tree(s) had a unique topology for all ten of the Betancur sets (albeit with an average of only 38 searches per set) as well as three additional sets (albeit only 115 and 256 searches, respectively, were performed for two of the Wiens sets). The optimal known tree was only present in 1–5% of all searches for 10 sets, and was only present in 6–10% of all searches for eight sets (note that matching topologies were not necessarily unique to sets of trees with the highest pseudo-likelihoods). The maximum scaled RF distance between the tree with the highest pseudo-likelihood and a suboptimal tree was 0 for Chiari DNA, 0.07 for Xi all characters, 0.08 for Chiari AA, 0.10 for Zhong, 0.13 for Pyron, 0.18 for Townsend, and 0.27 for McCormack.

3.8. Using the RF method to quantify improvements to Song and target additional outliers

The average scaled pairwise RF distance among Song’s original 447 gene trees for Mammalia is 0.52 (range 0.38–0.79), whereas the average distance for the 439 gene trees is 0.41 (range 0.29–0.79), and that for the 413 trees is 0.38 (range 0.28–0.77; Fig. 5). So, both stages of Springer and Gatesy’s (2016) improvements resulted in lower average scaled pairwise RF distances among the gene trees, but the more dramatic improvements came from using better fitting substitution models and running more thorough tree searches. This is evident when comparing the 447-gene-tree set to the 439-gene-tree set in Fig. 5, wherein the fraction of gene trees with average scaled pairwise RF distance >0.50 decreased substantially.

Twelve of the 26 gene trees eliminated by Springer and Gatesy (2016) because of sequence misalignment between different splice variants of the same gene had average scaled pairwise RF distance >0.50. The two outlier genes removed by Mirarab et al. (2014) had extreme average scaled pairwise RF of 0.76 and 0.79 and were among the 26 misaligned genes identified by Springer and Gatesy (2016). The 21 genes with swapped terminal taxa (a marsupial versus primate switch) that were eliminated by Mirarab et al. (2014) and fixed by Springer and Gatesy (2016) were characterized by high conflict with other gene trees. Prior to correction of the taxonomic swap, 18 of these genes had average scaled pairwise RF >0.5.

In addition to the outlier gene trees flagged by Springer and Gatesy (2016) and Mirarab et al. (2014), RF Distances Filter
revealed several additional gene trees with very high average scaled pairwise RF (>0.6). Sequence alignments for these nine conflicting gene trees indicate that eight are characterized by extensive missing data in multiple taxa and are short (average alignment length = 904 bp) relative to the mean for genes in the original Song dataset (3100 bp).

As an example, the ALOXSAP gene (arachidonate 5-lipoxygenase-activating protein; #203 from Song et al., 2012) includes only 486 aligned sites and has an average scaled pairwise RF distance of 0.625 (11th worst). Song et al.'s (2012) inferred gene tree implies numerous, inexplicable retentions of ancestral polymorphisms across branches that represent >96 million years (dos Reis et al., 2012). In addition to being short, there are extensive missing data at the 5' and 3' ends of this gene, with seven genera (Choloepus, Echinops, Erinaceus, Myotis, Sus, Tarsius, Tupaia) having 32–54% missing data. Because of the distribution of missing data at the 5' and 3' ends, some genera overlap for as few as only 70 sites (14%). Five ordinal-level taxa with extensive missing data are non-monophyletic (Cetartiodactyla, Chiroptera, Eulipotyphla, Primates, Xenarthra), as are the traditionally recognized mammalian clades Placantlia, Theria, Rodentia, and Glires.

### 4. Discussion

#### 4.1. Average pairwise distances among gene trees

The average scaled pairwise RF distances among gene trees varied greatly between studies (e.g., Wiens and Townsend vs. McCormack and Zhong) and in many cases within a given study (e.g., Chiari, Xi; Fig. 1). Studies and individual gene trees with high pairwise incongruence may alternatively be explained by the biological processes of lineage sorting, introgression, unrecognized sampling of paralogs, or the artifact of gene-tree reconstruction error. Gene-tree incongruence that cannot plausibly be caused by lineage sorting has already been described for McCormack (Gatesy and Springer, 2014), Zhong (Springer and Gatesy, 2014), and Xi (Simmons and Gatesy, 2015). Similarly, Betancur, Chiari, and

---

**Fig. 4.** Scaled Robinson-Foulds distance (RF dist.) between the optimal tree found by MP-EST and trees from all 1000 MP-EST searches (with the exception of Betancur and Wiens) for each of the 10 percentages of gene trees sampled for the eight empirical studies (A–M). The scaled RF distances are averaged over blocks of 50 trees. Subsamplings of gene trees were selected using the RF method.
Townsend pointed out the limited phylogenetic signal in some of their gene trees, and ascribed it to convergent nucleotide composition, saturation of third-codon positions, and lack of variation, respectively.

Based on the high variability in congruence among gene trees between studies and often within an individual study (Figs. 1 and 5), we recommend that presenting plots of pairwise RF distances among gene trees be treated as standard practice for empirical coalescent studies (e.g., Mirarab et al.’s (2014) Fig. S9, Pyron et al.’s (2014) Fig. 2). Distances among gene trees may also be effectively visualized using multidimensional scaling (Hillis et al., 2005; Pyron et al., 2014). Calculating average pairwise distances among gene trees while taking into account differential taxon sampling among gene trees is implemented in RF Distances Filter, which is freely available as supplemental online data at: http://dx.doi.org/10.6084/m9.figshare.1615928.

Furthermore, we assert that study authors should then examine gene trees along any long right-hand tail to evaluate whether their topologies are plausibly explained by lineage sorting. If they are not (based on, when available, the inferred times required for putative ancestral polymorphism to persist and coalescent-unit estimates; see Gatesy and Springer, 2014; Springer and Gatesy, 2014, 2016; Simmons and Gatesy, 2015) then the study authors should perform a sensitivity analysis (e.g., Wheeler, 1995) to test whether their inferred coalescent-based phylogeny is robust to exclusion of these highly incongruent gene trees. The favored phylogenetic hypothesis may be that which excludes misleading signal (Lecointre and Deleporte, 2005).

As an example of the above protocol, Springer and Gatesy (2016) reanalyzed Song by focusing on gene trees that were highly incongruent with the inferred species tree and then scrutinized DNA sequence alignments for these loci. Their ad hoc approach to examining gene-tree conflicts revealed several deficiencies in the dataset and analyses of Song; corrections of these problems resulted in striking improvements in congruence among gene trees (Fig. 5). However, the RF-method results presented here identified multiple additional gene trees in the long right-hand tail of the RF Distances Filter plot for Springer and Gatesy’s (2016) curated set of 413 improved gene trees (Fig. 5C). The additional problematic gene trees with pairwise RF > 0.6 are most simply explained by: (1) loci with extensive missing data for particular taxa that were identified by Springer and Gatesy (2016) but not removed from their 413 gene set, and (2) short loci (average of 904 bp) with missing data in multiple taxa.

In retrospect, an initial focus on gene-tree topologies with extreme pairwise scaled RF distances would have quickly identified many of the problems in the Song dataset, including editing errors, misalignment of different splice site variants, extensive missing data for certain taxa, and conflicting resolutions due to short locus length. Removal of all gene trees with pairwise RF > 0.5 from the curated set of 413 gene trees (39 gene trees removed) did not alter the coalescent topology that was consistently resolved by ASTRAL, MP-EST, and STAR in Springer and Gatesy (2016) wherein tree shrew (Scandentia) groups with Glirids (rodents and rabbit). This stable result conflicts with one of the main conclusions of Song (Scandentia + Primates clade), which was based on the 447-gene-tree set with much higher average pairwise RF (Fig. 5A).

4.2. ASTRAL vs. MP-EST vs. STAR

When comparing the performance of ASTRAL, MP-EST, and STAR with respect to the fractions of trees that are contrary to the reference clade(s) for each of the eight empirical studies, no method universally performed the best. But ASTRAL was the closest, generally outperforming MP-EST and/or STAR for seven of the eight studies (Fig. 2).

Likewise, no coalescent method universally outperformed the others with respect to congruence with the concatenation-based inferred phylogenies or internal consistency with the coalescent-based phylogenetic tree inferred from all gene trees (Figs. 3, S2). But ASTRAL was the closest, outperforming both MP-EST and STAR for the majority of the empirical studies sampled. The differences in average performance among the coalescent methods were often minor, but they are particularly apparent when comparing ASTRAL and STAR for the Betancur and Wiens studies, which include, by far, the greatest taxon sampling of the eight empirical studies sampled (214 and 171 terminals, respectively; Table 1) while including relatively few gene trees (20 and 44, respectively) with differential sampling of terminals. Hence these two studies, particularly when genes are subsampled, may be severe tests of the alternative coalescent methods.

Our results, wherein ASTRAL generally outperformed MP-EST and/or STAR, are consistent with simulation studies (Mirarab et al., 2014; Mirarab and Warnow, 2015) as well as Simmons and Gatesy’s (2015) contrived demonstration that MP-EST and STAR are more susceptible than ASTRAL to artifacts caused by misrooted gene trees. Based on our results and those of the earlier studies, we recommend that ASTRAL be used instead of MP-EST or STAR when coalescent analyses are applied to ancient lineages.

4.3. RF method

Subsampling gene trees using the RF method often, but certainly not always, performed better at resolving reference clades than did randomly subsampling gene trees (Fig. 2). But the more important comparison is how subsampling using the RF method performed relative to sampling all gene trees. In this context the utility of the RF method was more limited, and generally specific
to a single coalescent method for each of the applicable studies. Similarly, subsampling gene trees using the RF method was found to have limited utility with respect to improving congruence with the concatenation-based topology relative to the 100% gene-tree-sampling results or random subsamples of gene trees (Figs. S3 and S4).

We assert that subsampling gene trees using the RF method and then comparing the inferred topologies relative to the concatenation-based and 100% sampling-based inferred phylogenies is often a valuable data-exploration technique, but it cannot be generally expected to increase performance of coalescent methods relative to analyzing all gene trees (Fig. 2). We suggest, however, that the RF method is particularly well suited to testing whether gene-tree-estimation error is a primary cause of incongruence between concatenation-based and coalescent-based inferred phylogenies.

The hypothesis that gene-tree-estimation error had a more severe detrimental effect on coalescent rather than concatenation methods for a given study would be corroborated if the following three conditions are satisfied. First, many inferred gene trees are, on average, highly incongruent with other gene trees as determined based on pairwise RF distances. Second, subsampling gene trees using the RF method together with a coalescent method converges to the concatenation-based topology. Third, subsampling gene trees using the RF method converges to the concatenation-based topology more often than does using randomly subsampled gene trees. These three conditions may be checked using RF Distributions Filter. For example, these three conditions all apply to the gene trees. These three conditions may be checked using RF Distributions Filter. For example, these three conditions all apply to the gene trees.

Our null expectation is that gene-tree congruence is increased after eliminating artifacts (Doyle et al., 2015), and that choice among alternative character-coding and character-subsampling strategies may, in part, be determined using this criterion. With respect to the alternative character-coding strategies applied by Chiari, both DNA-based and AA-based gene trees have long right-hand tails of average pairwise RF distances, indicating that both include some highly incongruent gene trees (Fig. 1). But the DNA-based gene trees have higher average consistency with each other (0.51 for DNA vs. 0.70 for AA). Amino-acid characters are frequently treated as more conservative than nucleotide characters because they exclude silent substitutions. But there are other factors to consider as well, including reliance on fixed substitution-parameter values (Thorne et al., 1996), convergence (Simmons, 2000), and composite coding (Simmons and Freudenstein, 2002), which are specific to AA characters, and can result in discrepant clades (Simmons et al., 2002).

Despite Chiari’s DNA-based gene trees having higher average congruence with each other than did their AA-based gene trees (Fig. 1), coalescent analyses based on random subsamples of the DNA-based gene trees recovered all five of the reference clades far less often than did coalescent analyses based on random subsamples of the AA-based gene trees (Fig. 2). But note that subsampling the DNA-based gene trees using the RF method always recovered the reference clades when applied with all three coalescent methods. Based on these results we hypothesize that many of the DNA-based gene trees share a systematic bias (e.g., by long-branch attraction or convergent GC content), whereas the AA-based gene trees, albeit frequently incongruent with each other, perhaps simply because of stochastic error, do not share this systematic bias to the same degree.

With respect to the five alternative character-sampling strategies applied by Xi, the gene trees inferred using all characters had substantially lower pairwise incongruence than did the gene trees based on the OV-slow and TIGER-slow selected characters (all characters average: 0.56, OV-slow: 0.73, TIGER-slow: 0.73; Fig. 1). For Xi’s five character-sampling strategies we only have comparable data for resolution of the reference clades when all gene trees were sampled as well as when the RF method was applied to subsample gene trees (Figs. 2, 51). For all three coalescent methods, RF-based subsampling recovered both reference clades much more often when applied to the all-characters-based gene trees than the OV- or TIGER-selected character subsets. Taken together, these results reinforce Simmons and Gatesy’s (2015) conclusion that the OV- and TIGER-based subsampling strategies led to decreased phylogenetic signal when applied to Xi’s data, and the phylogenetic inferences based on sampling all of Xi’s characters should be preferred.
4.5. MP-EST tree searches

Authors of empirical studies who apply MP-EST often do not specify how many optimal-tree searches they performed (e.g., Chiari et al., 2012; Zhong et al., 2013; Xi et al., 2014; Stephens et al., 2015). Furthermore, some authors implement MP-EST using the STRAW web server (e.g., Tsagkogeorga et al., 2013; Cannon et al., 2015; Díaz-Rodríguez et al., 2015), but both STRAW servers only ever implement a single MP-EST search on uploaded data (L. Liu, pers. comm. 2015). Based on these two factors we expect that a large percentage of MP-EST analyses of empirical datasets are conducted using a single search.

Other than the Chiari datasets, which had by far the smallest taxon sample (16) of any of the sampled studies (Table 1), MP-EST rarely found the tree(s) with the highest known pseudo-likelihood in more than one search (Table S1). Different log likelihoods may be caused by rounding errors in likelihood calculations (e.g., Sanderson et al., 2015), different branch lengths for the same topology, and/or topological differences. Presumably most users of MP-EST primarily are interested in the topology rather than branch lengths, so we calculated scaled RF distances between the optimal tree found by MP-EST and trees from all MP-EST searches.

A single MP-EST search would have been sufficient to recover the optimal known tree topology for 28 of the 130 datasets (22%) for which subsamplings of gene trees were selected using the RF method (Fig. 4) and 24 of the 70 datasets (34%) for which subsamplings of gene trees were randomly selected (Fig. S5; note that the latter did not include any data from Betancur or Wiens because of computational limitations). Performing several MP-EST searches on a given dataset may also be insufficient to recover the optimal known tree topology, which was identified in ≤10% of all searches for 31 of the 130 datasets (24%) and 10 of the 70 datasets (14%), respectively. Differences in inferred MP-EST topologies were often not limited to a single branch swap among sister groups. At the extreme there was an RF distance of 14 (scaled RF distance = 0.27) between the tree with the highest pseudo-likelihood and the tree with the lowest pseudo-likelihood for a McCormack dataset. Likewise, there was an RF distance of 118 (scaled RF distance = 0.28) between the tree with the highest pseudo-likelihood and the tree with the lowest pseudo-likelihood for a Betancur dataset.

In the same manner that rigorous tree searches are important for traditional parsimony, likelihood, and Bayesian phylogenetic analyses (e.g., Goloboff, 1999; Hillis et al., 2005; Morrison, 2007), rigorous tree searches are also important for coalescent analyses. For datasets with more than a handful of taxa, we suggest that multiple tree-search replicates (preferably ≥100) should be implemented as a standard practice for empirical coalescent studies that apply MP-EST, at least in the context of using the default search-quality settings in MP-EST ver. 1.5. Liu et al. (2015) asserted that concatenation-based analyses overestimate bootstrap support because they do not take into account different gene-tree topologies. But another factor to consider is that low quality tree searches within MP-EST bootstrap pseudoreplicates may underestimate support because they fail to find the optimal tree topologies (see Freudenstein et al., 2004; Freudenstein and Davis, 2010).

5. Conclusions

Our five primary conclusions are as follows:

1. Average pairwise congruence among gene trees within a study varied greatly both between studies and also often within a study (Fig. 1).

2. Gene trees in long right-hand tails of average pairwise RF distances should be examined for artifacts that are not related to biological causes of gene-tree incongruence (Fig. 5).

3. Subsampling gene trees with high average pairwise RF distances may be used to reconcile conflicts between coalescent-based and concatenation-based phylogenetic inferences.

4. ASTRAL generally outperformed MP-EST and STAR with respect to resolving well supported reference clades (Fig. 2), self-consistency (Fig. S4), and congruence with concatenation results (Fig. 3), a pattern that also has emerged in recent simulation and empirical work at deep divergences.

5. Multiple tree-search replicates should be performed when using current default parameter settings for the coalescent method MP-EST (Table S1; Fig. 4).

Acknowledgments

We thank Ricardo Betancur-R and Edward Braun for many constructive criticisms with which to improve our manuscript; Ricardo Betancur-R, Frédéric Delsuc, Mark Springer, Ted Townsend, and John Wiens for sending trees; Sabrina Pankey for posting her Python script online; Toby Clarke for writing a Perl script to parse tree files; and Jessica Leigh for answering questions about Conclus-tador. D.B.S. was supported by NSF grant MCB-1412260. J.G. was supported by NSF grant DEB-1457735.

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.yjmpev.2015.12.013.

References

Crawford, N.G., Faircloth, B.C., McCormack, J.E., Brumfield, R.T., Winkler, K., Glenn, T.C., 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 8, 79–786.