Molecular Phylogenetics and Evolution 92 (2015) 63–71

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased ☆

Zhenxiang Xi [a], Liang Liu [b,c], Charles C. Davis [a,*]

[a] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA
[b] Department of Statistics, University of Georgia, Athens, GA 30602, USA
[c] Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA

ARTICLE INFO

Article history:
Received 27 January 2015
Revised 23 April 2015
Accepted 16 June 2015
Available online 24 June 2015

Keywords:
Concatenation methods
Gene informativeness
Gene tree estimation
Multilocus bootstrap approach
Gene-tree-based coalescent methods
PhyML

ABSTRACT

The development and application of coalescent methods are undergoing rapid changes. One little explored area that bears on the application of gene-tree-based coalescent methods to species tree estimation is gene informativeness. Here, we investigate the accuracy of these coalescent methods when genes have minimal phylogenetic information, including the implementation of the multilocus bootstrap approach. Using simulated DNA sequences, we demonstrate that genes with minimal phylogenetic information can produce unreliable gene trees (i.e., high error in gene tree estimation), which may in turn reduce the accuracy of species tree estimation using gene-tree-based coalescent methods. We demonstrate that this problem can be alleviated by sampling more genes, as is commonly done in large-scale phylogenomic analyses. This applies even when these genes are minimally informative. If gene tree estimation is biased, however, gene-tree-based coalescent analyses will produce inconsistent results, which cannot be remedied by increasing the number of genes. In this case, it is not the gene-tree-based coalescent methods that are flawed, but rather the input data (i.e., estimated gene trees). Along these lines, the commonly used program PhyML has a tendency to infer one particular bifurcating topology even though it is best represented as a polytomy. We additionally corroborate these findings by analyzing the 183-locus mammal data set assembled by McCormack et al. (2012) using ultra-conserved elements (UCEs) and flanking DNA. Lastly, we demonstrate that when employing the multilocus bootstrap approach on this 183-locus data set, there is no strong conflict between species trees estimated from concatenation and gene-tree-based coalescent analyses, as has been previously suggested by Gatesy and Springer (2014).

1. Introduction

Advances in next-generation sequencing and computational phylogenomics have shifted the emphasis of phylogenetic studies from gene tree to species tree estimation (Edwards, 2009). Until recently, the reconstruction of phylogenies using genomic data has relied largely on standard concatenation methods, in which phylogenies are inferred from a single matrix assembled by concatenating hundreds of genes (William and Ballard, 1996; de Queiroz and Gatesy, 2007). In contrast to concatenation methods, which implicitly assume that all genes have the same or very similar evolutionary histories, recently developed coalescent-based methods instead permit gene trees to have different evolutionary histories (Liu et al., 2009a). The application of concatenation versus coalescent methods to deep phylogenetic questions, however, has become controversial in recent years (Gatesy and Springer, 2013, 2014; Wu et al., 2013; Springer and Gatesy, 2014; Zhong et al., 2014; Liu et al., 2015b). In many cases, phylogenomic analyses applying these methods have converged on similar topologies, but in others, concatenation and coalescent methods have yielded conflicting relationships (e.g., Song et al., 2012; Xi et al., 2013, 2014).

Under what circumstances do conflicts between concatenation and coalescent analyses arise? And when might one method be preferred? Simulation studies have shown that when the degree of incomplete lineage sorting (ILS) is low or gene tree estimation error is relatively high, concatenation methods can be more accurate than coalescent methods (Mirarab et al., 2014, 2015). At the same time, recent simulation and empirical studies have demonstrated that compared to concatenation methods, coalescent

☆ This paper was edited by the Associate Editor Derek Wildman.

* Corresponding author.
E-mail address: cdavis@oeb.harvard.edu (C.C. Davis).

http://dx.doi.org/10.1016/j.ympev.2015.06.009
1055-7903/

methods better accommodate gene tree discordance due to ILS (Liu et al., 2008, 2009b, 2010; Kubatko et al., 2009; Leaché and Rannala, 2011; Song et al., 2012; Zhong et al., 2013; Mirarab et al., 2014, 2015), and are more robust to elevated nucleotide substitution rates (Xi et al., 2013, 2014; Liu et al., 2015a). Importantly, in a recent review on this topic, Liu et al. (2015b) have demonstrated that the concatenation model, with or without partitions, is a simplified case of the multispecies coalescent model. This occurs when gene tree topologies are identical to each other and ultimately identical to the species tree.

The development and application of coalescent methods are undergoing rapid changes. Some of these methods, such as BEST (Liu, 2008) and *BEAST (Heled and Drummond, 2010), simultaneously estimate the gene trees and species tree. These co-estimation methods have outstanding accuracy, but are computationally intensive and do not scale up for genome-level analyses (Leaché and Rannala, 2011; Bayzid and Warnow, 2013; Mirarab et al., 2015). Thus, they are not the focus of our study here. Instead, we focus on gene-tree-based coalescent methods, which infer the species tree from a set of estimated gene trees as implemented in MP-EST (Liu et al., 2010), STELLS (Wu, 2012), and STEM (Kubatko et al., 2009). In addition, some of the recently developed consensus methods, such as ASTRAL (Mirarab et al., 2014), NJ_{st} (Liu and Yu, 2011), STAR (Liu et al., 2009b), and STEAC (Liu et al., 2009b), estimate the species tree using summary statistics from the estimated gene trees. Although the latter consensus methods are not strictly coalescent-based, they can accommodate gene tree discordance due to ILS, and have been shown to be statistically consistent under the multispecies coalescent model (Liu et al., 2009b; Liu and Yu, 2011; Mirarab et al., 2014). For simplicity, we also refer to these consensus methods as gene-tree-based coalescent methods.

One little explored area that bears on the application of gene-tree-based coalescent methods to phylogenomic data is gene informativeness. In a recent study seeking to resolve relationships among placental mammals (i.e., Placentalia), McCormack et al. (2012) assembled a 183-locus data set using ultra-conserved elements (UCEs) and flanking DNA across 29 species (27 mammals, one bird, and one reptile). UCEs are DNA sequences that are identical, or nearly so, and are located in syntenic positions in at least two genomes (Reneker et al., 2012). Gatesy and Springer (2014) subsequently used this complete data set containing no missing genes to evaluate the performance of standard concatenation versus gene-tree-based coalescent methods (termed "shortcut" coalescent methods by Gatesy and Springer (2014)). In their analyses, the species trees inferred from coalescent analyses (MP-EST and STAR) using gene trees estimated by PhyML (Guindon et al., 2010) conflicted with species trees inferred from concatenation analyses using GARLI (Zwickl, 2006) and RAxML (Stamatakis, 2014). The authors noted that this incongruence was especially notable within the Glires clade, which included lagomorphs and rodents. Based on these conflicting results, the authors concluded that inaccurate reconstruction of gene trees was especially problematic for gene-tree-based coalescence methods.

One concern that arises from these conclusions by Gatesy and Springer (2014) is that they have unnecessarily confounded gene tree estimation with the general premise of gene-tree-based coalescent methods. In particular, we noticed that many of these 183 UCE loci are short (ranging from 103 to 1036 nucleotide sites, with a mean of 408 sites per locus) and the number of parsimony informative sites within each locus is low (ranging from 2 to 264, with a mean of 44 per locus). This renders many loci potentially lacking phylogenetic information for gene tree estimation. These numbers are striking when compared to other phylogenomic studies of placental mammals, including the 26-gene data set analyzed by Meredith et al. (2011) (a mean of 1369 nucleotide sites and 933 parsimony informative sites per gene), the 447-gene data set analyzed by Song et al. (2012) (a mean of 3101 nucleotide sites and 1196 parsimony informative sites per gene), and the 857-gene data set analyzed by dos Reis et al. (2012) (a mean of 2955 nucleotide sites and 1016 parsimony informative sites per gene). Such a lack of phylogenetic information is likely to be especially problematic for regions of the species tree where internal branches are very short (Townsend, 2007). Under these circumstances, gene trees estimated from alignments with minimal phylogenetic information may reduce the accuracy of gene-tree-based coalescent methods (or any coalescent method for that matter), in the same way that weak or uninformative concatenated alignments will reduce the accuracy of concatenation methods. We suspect that as long as gene tree estimation is not biased, the accuracy of gene-tree-based coalescent methods can be improved by sampling more genes, even when these genes are minimally informative. This area, however, remains largely unexplored and will be increasingly important in the phylogenomic era.

We addressed this issue from two different perspectives. First, we utilized simulated DNA sequences to assess whether gene alignments with minimal phylogenetic information could compromise gene-tree-based coalescent analyses. Second, we reanalyzed the 183-locus data set using concatenation and gene-tree-based coalescent methods. In the same study by Gatesy and Springer (2014), no statistical confidence was assessed for their species trees inferred from the 183-locus data set using gene-tree-based coalescent methods. Thus, it is not possible to determine if the conflicts they identified in concatenation versus gene-tree-based coalescent analyses are well supported. Here, we consider clades with bootstrap percentage (BP) support ranging from 80 to 100 to be well supported. For the concatenation analyses performed by Gatesy and Springer (2014), statistical confidence was estimated using the conventional non-parametric bootstrap approaches for the combined, non-partitioned matrix (i.e., standard (Felsenstein, 1985) and rapid (Stamatakis et al., 2008) bootstrap approaches). These conventional bootstrap approaches resample sites from concatenated sequences with replacement, and assume that sites are independently and identically distributed. This assumption, however, is inappropriate for large-scale phylogenomic data since different genes have been subject to different evolutionary processes (Seo, 2008). In addition, for conventional bootstrap procedures, the resulting tree will reflect mainly the tree supported by long genes and their gene-specific features (Seo et al., 2005). In contrast, recent studies have shown that the multilocus bootstrap approach, in which genes are resampled with replacement followed by resampling sites with replacement within each gene, can more effectively account for variation in gene-specific evolutionary features and reduce the effect of gene length (Seo et al., 2005; Seo, 2008). We assessed statistical confidence for the 183-locus data set using the multilocus bootstrap approach, and compared the results between the multilocus and conventional bootstrap analyses.

2. Material and methods

2.1. DNA simulations to investigate the effects of gene length and the number of phylogenetically informative sites on gene-tree-based coalescent methods

We simulated DNA sequences on eight ultrametric species trees under a variety of branch lengths (to simulate rate variation) and population size parameters (to simulate various degrees of ILS) using the multispecies coalescent model (Rannala and Yang, 2003). These 5-taxon species trees, T1–T8, are topologically identical except with respect to branch lengths and the degree of ILS

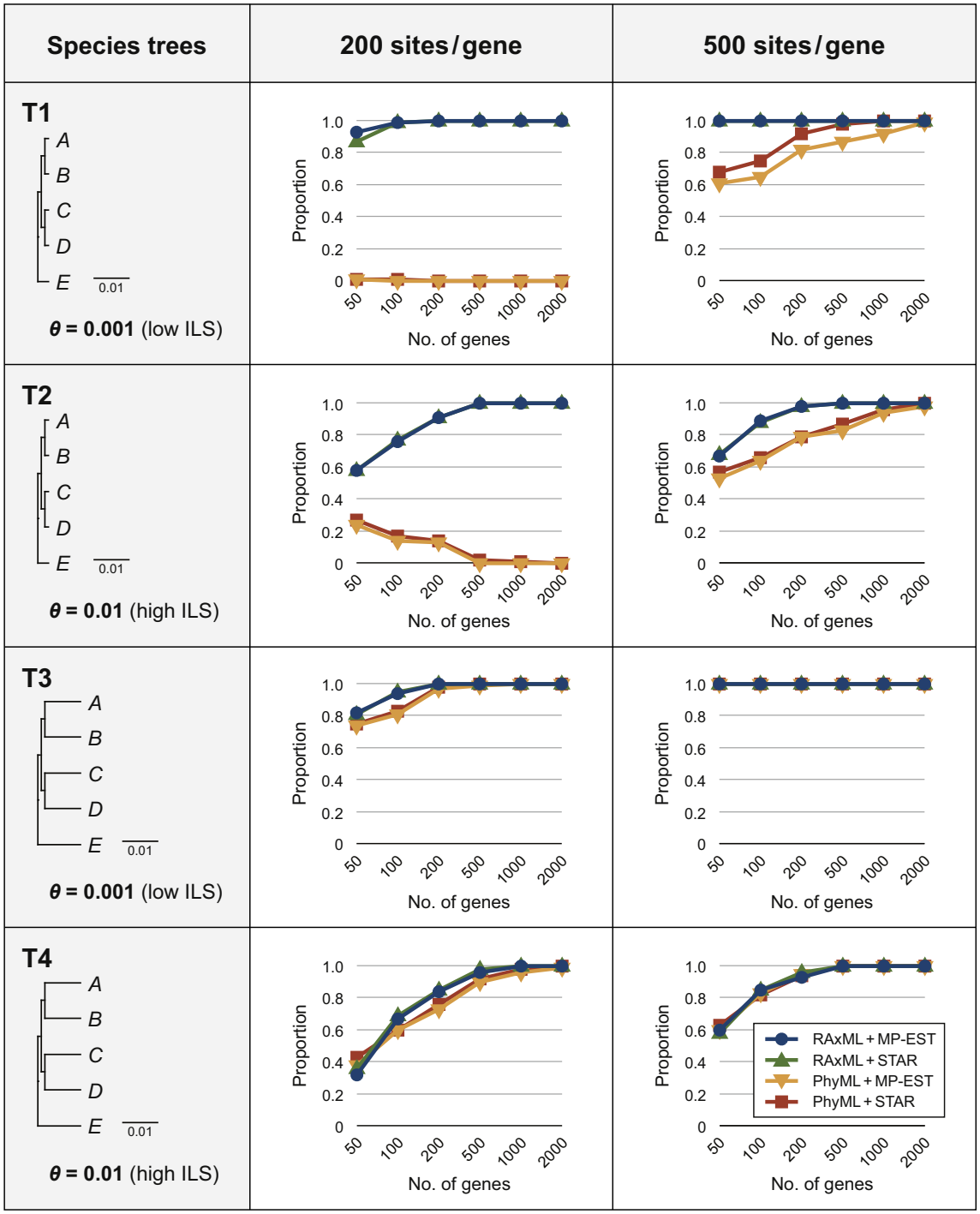

Fig. 1. DNA simulations to investigate the effects of gene length and the number of phylogenetically informative sites on gene-tree-based coalescent methods (MP-EST and STAR). DNA sequences were simulated on eight ultrametric species trees T1–T8 under the multispecies coalescent model (Rannala and Yang, 2003). The lengths of the three internal branches are 0.001 and 0.01 for species trees T1–T4 and T5–T8, respectively (branch lengths are in mutation units, i.e., the number of nucleotide substitutions per site). The lengths of the external branches leading to ingroup species *A*–*D* are 0.001 for species trees T1, T2, T5, and T6, and 0.01 for species trees T3, T4, T7, and T8. In addition, we applied three different values of the population size parameter θ to simulate various degrees of incomplete lineage sorting (ILS). The population size parameter θ is defined as $4\mu N_e$, where N_e is the effective population size and μ is the average mutation rate per site per generation. Individual gene trees were estimated using PhyML and RAxML. Results shown here represent the proportions of simulations in which MP-EST and STAR recover the true species tree.

(Fig. 1). In each of these species trees, species *E* was designated as the outgroup. For each gene, one allele was sampled from each of the species *A*–*E*. We varied the lengths of the three internal branches across these eight species trees (Fig. 1). For species trees T1–T4, internal branches are short (i.e., 0.001; branch lengths are in mutation units, i.e., the number of nucleotide substitutions per site); whereas for species trees T5–T8, they are long (i.e., 0.01). We also varied the lengths of the external branches across these species trees (Fig. 1). For species trees T1, T2, T5, and T6, the lengths of the external branches leading to ingroup species *A*–*D* are short (i.e., 0.001); whereas for species trees T3, T4, T7, and T8, they are long (i.e., 0.01).

In addition, we applied three different values of the population size parameter θ to simulate varying degrees of ILS (i.e., 0.001 for

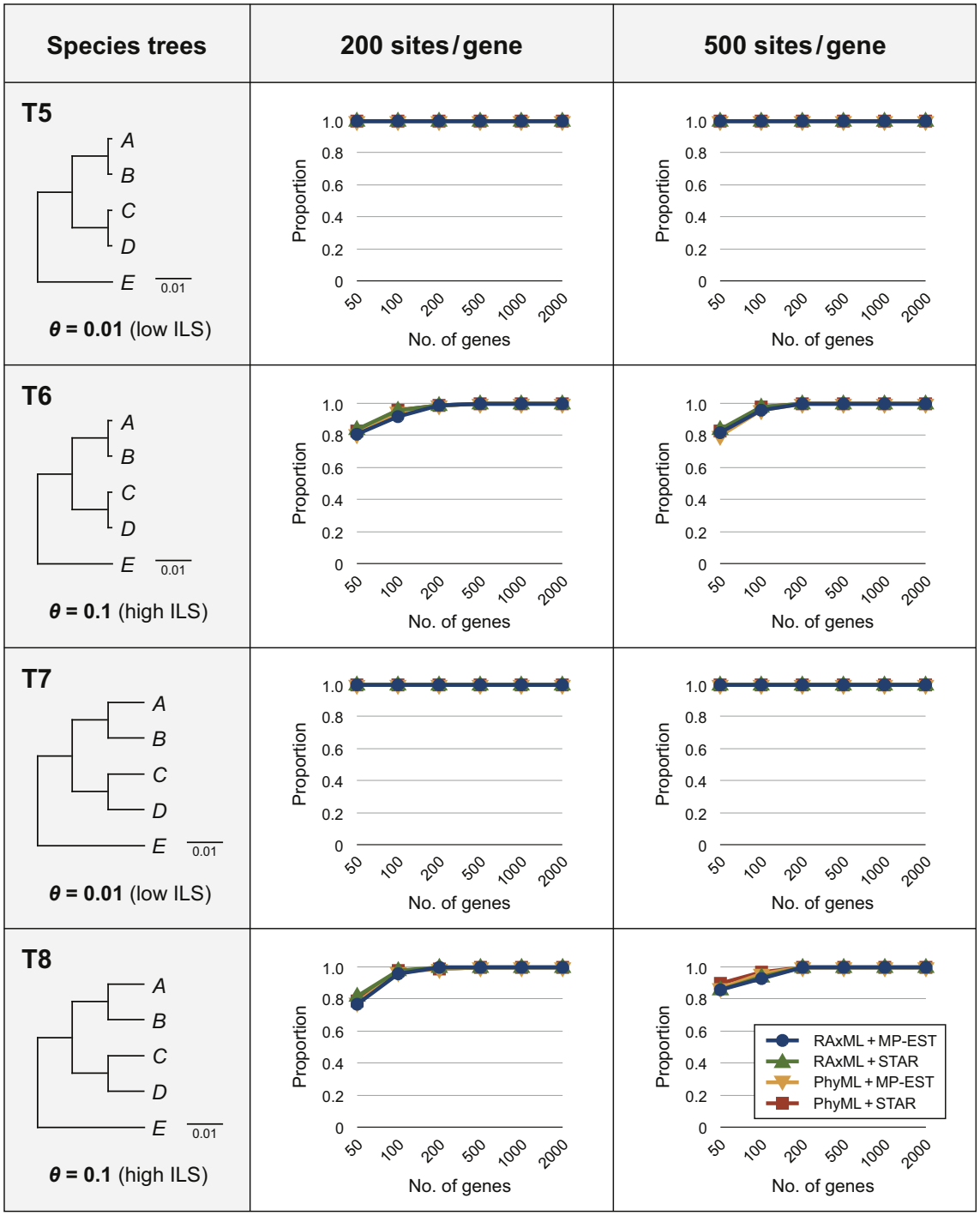

Fig. 1 (*continued*)

species trees T1 and T3, 0.01 for species trees T2, T4, T5, and T7, and 0.1 for species trees T6 and T8; Fig. 1). These assigned values of θ were selected based on previous simulation studies (Liu et al., 2010; Xi et al., 2014), and are broadly consistent with empirical values of θ for many animals and plants (Degnan and Rosenberg, 2009). The population size parameter θ is defined as $4\mu N_e$, where N_e is the effective population size and μ is the average mutation rate per site per generation. For each of these species trees, we assumed that population size was constant across all populations. According to coalescent theory, the degree of ILS is negatively correlated with branch lengths on the species tree, as measured in coalescent units (i.e., branch lengths in mutation units divided by θ). Thus, gene trees simulated on species trees with short internal branches (in coalescent units) are highly discordant regarding their topologies. In this regard, for species trees T1, T3, T5, and T7, the degree of ILS is low (i.e., 1.0 coalescent units); while for species T2, T4, T6, and T8, the degree of ILS is high (i.e., 0.1 coalescent units).

Next, we simulated 50, 100, 200, 500, 1000, and 2000 gene trees on each of species trees T1–T8 using the R function *sim.coaltree.sp* as implemented in Phybase v1.3 (Liu and Yu, 2010). Each gene tree was then utilized to simulate one DNA alignment using Seq-Gen

v1.3.3 (Rambaut and Grassly, 1997) with the JC69 model (Jukes and Cantor, 1969). To evaluate how the length of gene alignments affected species tree estimation, we simulated gene alignments with two length categories (i.e., 200 and 500 nucleotide sites; Fig. 1). The lengths of these simulated gene alignments are comparable with the average locus length of the 183-locus data set.

For gene-tree-based coalescent analyses, individual gene trees were first inferred from RAxML v8.1.3 using the GTR+Γ model with a random starting tree ("-d -f o -m GTRGAMMAX --no-bfgs --no-seq-check"), and rooted with species *E*. These estimated gene trees were then utilized to construct the species trees using MP-EST v1.4 and the STAR method as implemented in Phybase (default settings were used for MP-EST and STAR). Each simulation was repeated 100 times. To compare gene tree estimation using different maximum likelihood (ML) programs, we additionally inferred individual gene trees using the GTR+Γ model with a random starting tree in GARLI v2.01.1067 ("ratematrix = 6rate statefrequencies = estimate ratehetmodel = gamma numratecats = 4 invariantsites = none streefname = random searchreps = 1 collapsebranches = 0") and PhyML v20120412 ("-a e -b 0 -c 4 -f m -m GTR -o tlr -s SPR -v 0 --rand_start --n_rand_starts 2").

2.2. The 183-locus mammal data set

We reanalyzed the 183-locus data set assembled by McCormack et al. (2012) using both concatenation and gene-tree-based coalescent methods. For concatenation analyses, the best-scoring ML trees were estimated from the concatenated nucleotide matrix using a single GTR+Γ model following Gatesy and Springer (2014). Analyses were performed with a random starting tree using PhyML ("-a e -b 0 -c 4 -f m -m GTR -o tlr -s SPR -v 0 --rand_start --n_rand_starts 2") and RAxML ("-d -f o -m GTRGAMMAX --no-bfgs --no-seq-check"). For gene-tree-based coalescent analyses, individual gene trees were first inferred using PhyML and RAxML as described in Section 2.1, and rooted with anole (*Anolis carolinensis*). These estimated gene trees were then utilized to construct the species tree using MP-EST and STAR. Bootstrap support was estimated for concatenation and gene-tree-based coalescent methods using the multilocus bootstrap approach (Seo, 2008) with 100 replicates. For comparison, we additionally estimated bootstrap support from the concatenated 183-locus matrix using the rapid bootstrapping as implemented in RAxML ("-d -f a -m GTRGAMMAX -N 100 --no-seq-check --no-bfgs"). For standard bootstrapping, we first created 100 bootstrapped matrices from the original concatenated matrix ("-f j -m GTRGAMMAX -N 100 --no-seq-check --no-bfgs"), and then inferred the ML trees from bootstrapped matrices using RAxML as described above.

3. Results and discussion

3.1. The effects of gene length and the number of phylogenetically informative sites on gene-tree-based coalescent methods

Simulation analyses of species trees T1–T8 (Fig. 1) demonstrated that when the degree of ILS was low (i.e., species trees T1, T3, T5, and T7), on average 83% of the simulated gene trees (when rooted with species *E*) were congruent with the species tree topology (Fig. 2A). When the degree of ILS was high (i.e., species trees T2, T4, T6, and T8), the topologies of simulated gene trees were highly variable despite a common species tree. Under these circumstances, on average only 20% of the simulated gene trees were congruent with the species tree topology (Fig. 2A).

When the lengths of the three internal branches were long (i.e., 0.01 in species trees T5–T8; Fig. 1), both MP-EST and STAR accurately estimated the true species trees as the number of genes increased. Here, regardless of the ML program used for gene tree estimation (i.e., GARLI, PhyML, or RAxML), the proportion of simulations in which MP-EST and STAR recover the true species tree increased to $\geqslant$0.99 as the number of genes increased to 200. In these cases, the percentage of estimated gene trees that matched the species tree topology was high for both short (i.e., 200 sites) and longer (i.e., 500 sites) genes: on average more than 63% and 18% of estimated gene trees matched the species tree topology when the degree of ILS was low and high, respectively (Fig. 2B), compared to on average 83% and 20% of simulated gene trees that matched the species tree topology when the degree of ILS was low and high, respectively (Fig. 2A). These comparable results from simulated and estimated gene trees establish that as long as sufficient phylogenetic information is present in each gene (as simulated here using the longer internal branches in the species tree), the performance of gene-tree-based coalescent methods are not adversely affected by short genes.

When the lengths of the three internal branches were short (i.e., 0.001 in species trees T1–T4; Fig. 1), however, the distribution of estimated gene trees differed dramatically from the distribution of simulated gene trees, especially when genes were short (i.e., 200 sites). For example, despite that on average 83% of the simulated gene trees were congruent with the species tree T3 (Fig. 2A), on average only 16% of the gene trees estimated by RAxML matched the species tree topology (Fig. 2B). In these short genes, the number of parsimony informative sites was on average less than one per gene for species trees T1 and T3, compared to on average four sites per gene for species trees T5 and T7. Despite these differences, the most frequent gene tree estimated by GARLI and RAxML from short genes still matched the species tree topology (e.g., on average 22% and 13% for species trees T1 and T2, respectively, using RAxML; Fig. 2B). Thus, in these cases, coalescent analyses using gene trees estimated by GARLI and RAxML (but not PhyML, see below) still recovered the true species tree with a proportion of 1.0 as the number of genes increased to 500 (Fig. 1). These results indicate that genes with minimal phylogenetic information can greatly reduce the accuracy of gene tree estimation, however, the performance of gene-tree-based coalescent methods can be greatly improved by sampling more genes, even if they are relatively short and minimally informative.

In contrast, the most frequent gene tree (i.e., topology II in Fig. 2A) estimated by PhyML from these short genes (i.e., 200 sites) did not match species trees T1 and T2 (Fig. 2C). Under these circumstances, gene tree estimation was biased toward one particular incorrect gene tree even though a random starting tree was used for heuristic tree searching. For example, although this incorrect gene tree constituted on average only 4% of all simulated gene trees for the species tree T1 (Fig. 2A), it constituted on average 45% of all estimated gene trees (Fig. 2C). Here, the proportion of simulations in which MP-EST and STAR recovered the true species tree decreased to $\leqslant$0.02 as the number of genes increased to 50 and 500 for species trees T1 and T2, respectively (Fig. 1). In these cases, MP-EST and STAR consistently inferred the incorrect species tree (i.e., topology II in Fig. 2A) with a proportion of 1.0 as the number of genes increased. Furthermore, when increasing the number of sites per gene (i.e., 500 sites), and hence increasing the number of parsimony informative sites, coalescent analyses using gene trees estimated by PhyML recovered the true species tree with a proportion of $\geqslant$0.98 as the number of genes increased to 2000 for species trees T1 and T2 (Fig. 1). Therefore, in these cases, the performance of coalescent methods appears to be greatly compromised by highly biased gene tree estimation using PhyML when individual genes have minimal phylogenetic information. This assertion is further supported by the fact that this bias is greatly reduced when using genes that are more informative.

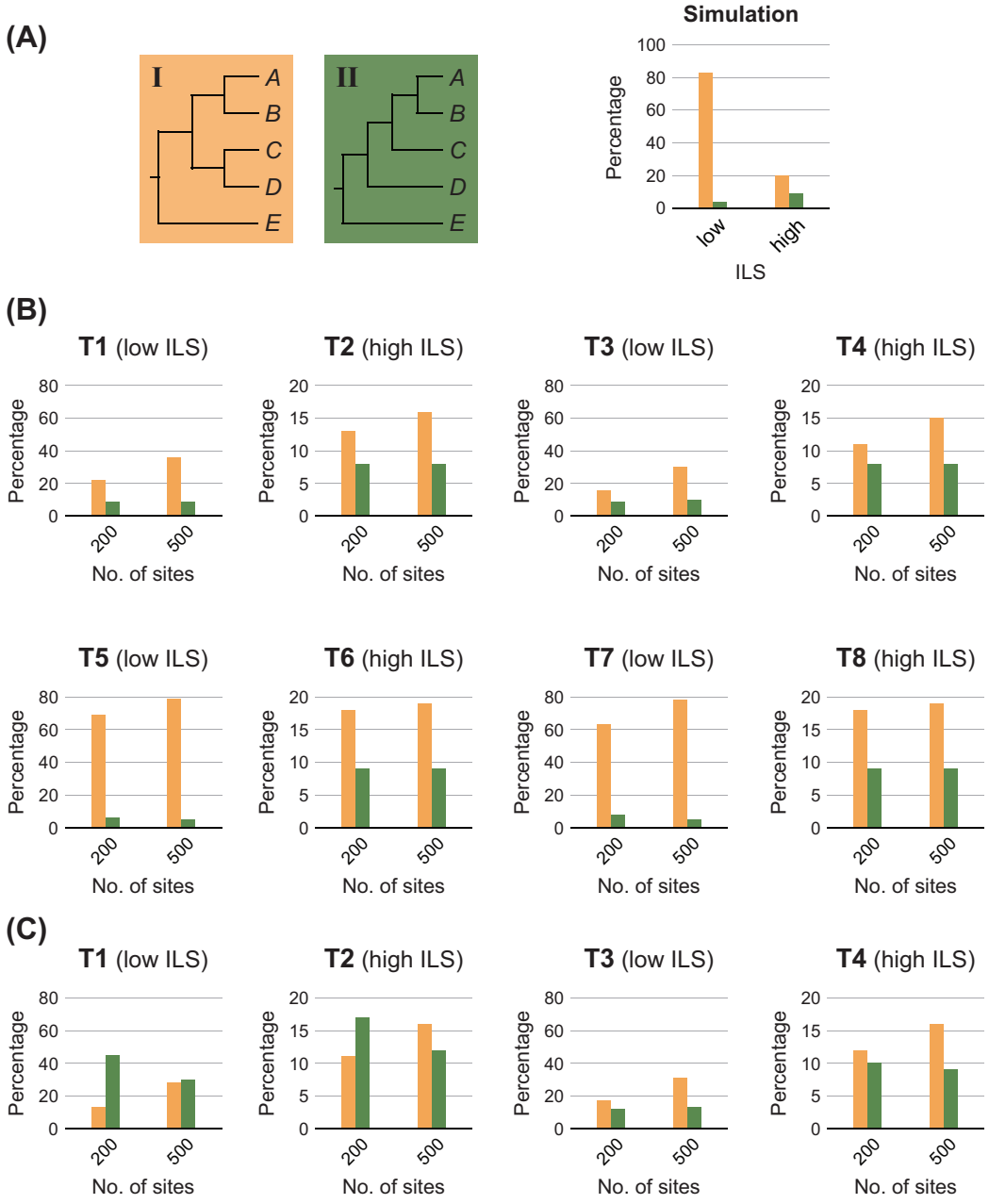

Fig. 2. The frequency spectrum of topologies I (orange) and II (green) in simulated and estimated gene trees. (A) The average percentages of the two topologies in simulated gene trees (rooted with species *E*) when the degree of incomplete lineage sorting (ILS) is low or high. (B) The average percentages of the two topologies in gene trees estimated by RAxML. (C) The average percentages of the two topologies in gene trees estimated by PhyML. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

What may explain this bias in gene tree estimation? Our simulation suggests that gene trees estimated from genes with minimal phylogenetic information can dramatically differ from the actual gene trees, as has been shown elsewhere (e.g., Hillis et al., 1994; Rasmussen and Kellis, 2007; Townsend, 2007; Capella-Gutierrez et al., 2014). Since many commonly used phylogenetic programs, such as PhyML, produce strictly bifurcating trees, gene trees estimated from alignments with minimal phylogenetic information likely possess artifactual relationships that should be best represented as unresolved, i.e., polytomies. RAxML similarly resolves trees as bifurcated, but unlike PhyML, our simulation analyses demonstrate that RAxML does not appear to be biased toward one particular topology under these circumstances. Along these lines, we suspect that gene trees that were initially estimated by McCormack et al. (2012) using PhyML, and subsequently adopted by Gatesy and Springer (2014), likely contributed to some of Gatesy and Springer's conclusions on the downside of gene-tree-based coalescent methods (see also Section 3.2 below). While we agree that this finding highlights the importance of gene tree estimation, we disagree with their claim that this is a critical flaw of gene-tree-based coalescent methods more generally. Like traditional phylogenetic methods, gene-tree-based coalescent methods will not perform well when the input data (i.e., estimated gene trees) are biased as appears to have been the case here.

Similar results have been reported in a recent simulation study (Mirarab et al., 2015), demonstrating that gene tree estimation error is high for short genes (which can be less informative than longer genes), especially when species tree branch lengths are

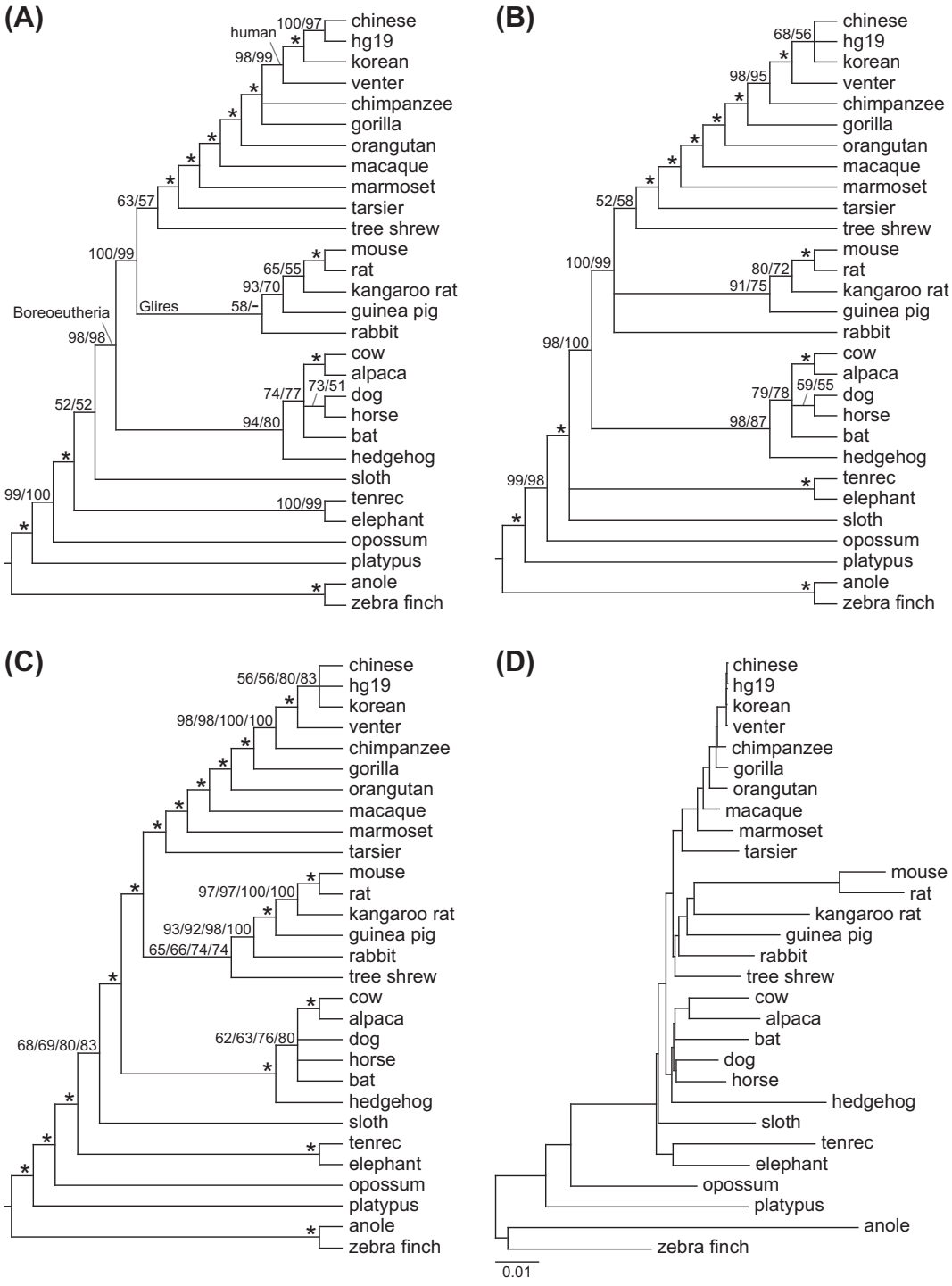

Fig. 3. The species trees inferred from the 183-locus mammal data set using gene-tree-based coalescent (MP-EST and STAR) and concatenation (PhyML and RAxML) methods. These 183 loci were originally assembled by McCormack et al. (2012) using ultra-conserved elements (UCEs) and flanking DNA. (A) The 50% majority-rule multilocus bootstrap consensus trees inferred from MP-EST and STAR using gene trees estimated by PhyML. Bootstrap percentages (BPs) estimated by STAR/MP-EST using the multilocus bootstrapping are indicated for each branch. An asterisk indicates that the branch is supported by 100 BPs. (B) The 50% majority-rule multilocus bootstrap consensus trees inferred from MP-EST and STAR using gene trees estimated by RAxML. BPs estimated by STAR/MP-EST using the multilocus bootstrapping are indicated for each branch. An asterisk indicates that the branch is supported by 100 BPs. (C) The 50% majority-rule multilocus bootstrap consensus trees inferred from PhyML and RAxML. BPs estimated by PhyML using the multilocus bootstrapping/RAxML using the multilocus bootstrapping/RAxML using the standard bootstrapping/RAxML using the rapid bootstrapping are indicated for each branch. An asterisk indicates that the branch is supported by 100 BPs. (D) The best-scoring maximum likelihood tree inferred from RAxML. Branch lengths are in mutation units, i.e., the number of nucleotide substitutions per site.

short. With regard to empirical data, if estimated gene trees differ from the actual gene trees due to random error (e.g., gene trees estimated by RAxML in our simulation analyses), the performance of gene-tree-based coalescent methods can be improved by sampling more genes (even if they are minimally informative). However, if gene tree estimation is biased toward one particular incorrect topology (e.g., gene trees estimated by PhyML in our simulation analyses) or significantly affected by systematic errors (e.g., substitution saturation at third codon positions (Chiari et al., 2012)), gene-tree-based coalescent methods will produce

inconsistent results, which cannot be remedied by increasing the number of genes. In this case, it is not the species tree estimation method that is flawed, but rather the input data (i.e., estimated gene trees for gene-tree-based coalescent methods). In this regard, our simulation analyses indicate that when genes are short or less informative, GARLI and RAxML are preferred to PhyML in relation to gene tree estimation.

3.2. The 183-locus mammal data set

Our gene-tree-based coalescent analyses (MP-EST and STAR) of the 183-locus data set produced generally well resolved species trees using the multilocus bootstrap approach (Fig. 3A and B). Despite the fact that there was a lack of phylogenetically informative sites for many of the 183 loci, the majority of relationships were supported with >80 BP. Based on our simulation analyses, these empirical results further indicate that even when individual genes are less informative, the performance of gene-tree-based coalescent methods can be improved by sampling a large number of genes. In addition, relationships among the four human accessions labeled "chinese", "hg19", "korean", and "venter" were poorly resolved ($\leqslant$68 BP) in our coalescent analyses using gene trees estimated by RAxML (Fig. 3B). In contrast, these relationships were strongly supported ($\geqslant$97 BP) using gene trees estimated by PhyML (Fig. 3A). The branch lengths associated with these four human accessions were especially short (Fig. 3D) and very reminiscent of species trees T1 and T2 as described in Section 3.1. We suspect that the highly inflated BPs recovered in coalescent analyses using gene trees estimated by PhyML are artifactual and attribute to its bias in gene tree estimation as described in Section 3.1. Collectively, our simulation and empirical analyses suggest genes with minimal phylogenetic information can produce biased gene tree estimation using PhyML, which can result in inconsistent results when using gene-tree-based coalescent methods.

Lastly, the multilocus bootstrap consensus trees inferred from coalescent analyses using gene trees estimated by RAxML (Fig. 3B) and concatenation analyses (PhyML and RAxML; Fig. 3C) of the 183-locus data set largely agreed with each other. There were no strongly conflicting phylogenetic relationships (i.e., >80 BP) between our concatenation and gene-tree-based coalescent analyses. We identified only one relationship (i.e., the monophyly of Glires) that was well supported by concatenation methods but unresolved (i.e., <50 BP) in gene-tree-based coalescent analyses. Thus, the conclusion by Gatesy and Springer (2014) that gene-tree-based coalescent methods produced "contradictory" results for this data set does not apply after properly assessing statistical confidence. In addition, given the general congruence between these species trees (i.e., all well-supported relationships identified in our gene-tree-based coalescent analyses were also supported with >80 BP in our concatenation analyses; Fig. 3B and C), the statement by Gatesy and Springer (2014) that support for incorrect relationships using gene-tree-based coalescent methods was the result of artifacts is also not supported by our analyses.

Furthermore, our analyses indicate that the conventional bootstrap analyses produce overall higher BPs than the multilocus bootstrap approach. For example, the standard and rapid bootstrap analyses supported the placement of sloth as sister to Boreoeutheria and the monophyly of the three human accessions labeled "chinese", "hg19", and "korean" with $\geqslant$80 BP (Fig. 3C). In contrast, support for this relationship was much lower using the multilocus bootstrap approach (<70 BP by PhyML and RAxML; Fig. 3C). In these cases, well-supported relationships from conventional bootstrap analyses are shown to be poorly supported upon employing the multilocus bootstrap approach. Given that the latter approach has been demonstrated to more effectively account for gene-specific evolutionary features and reduce the effect of gene length (Seo et al., 2005; Seo, 2008), we recommend its usage for large-scale phylogenomic analyses of this nature.

4. Conclusion

Our simulation and empirical analyses collectively demonstrate that genes with minimal phylogenetic information can produce unreliable gene trees (i.e., high error in gene tree estimation), which may in turn reduce the accuracy of species tree estimation using gene-tree-based coalescent methods. We demonstrate that this problem can be alleviated by sampling more genes, as is commonly done in large-scale phylogenomic analyses. This applies even when these genes are minimally informative. If gene tree estimation is biased, however, gene-tree-based coalescent analyses will produce inconsistent results, which cannot be remedied by increasing the number of genes. In this case, it is not the gene-tree-based coalescent methods that are flawed, but rather the input data (i.e., estimated gene trees). Along these lines, the commonly used program PhyML has a tendency to infer one particular bifurcating topology even though it is best represented as a polytomy. This problem can be remedied by sampling genes with increased phylogenetic information or using a different program for gene tree estimation (e.g., RAxML).

Acknowledgments

We thank Jeffrey Dacosta, Scott Edwards, members of the Davis and Liu laboratories, and the Harvard Phylogenetics Journal Club for helpful comments and discussion. We also thank Andre Aberer and Stéphane Guindon for technical support. This work was supported by the United States National Science Foundation (DEB-1120243 to C.C.D. and DMS-1222745 to L.L.).

References

Bayzid, M.S., Warnow, T., 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29, 2277–2284.
Capella-Gutierrez, S., Kauff, F., Gabaldón, T., 2014. A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucleic Acids Res. 42, e54.
Chiari, Y., Cahais, V., Galtier, N., Delsuc, F., 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10, 65.
de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41.
Degnan, J.H., Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24, 332–340.
dos Reis, M., Inoue, J., Hasegawa, M., Asher, R.J., Donoghue, P.C.J., Yang, Z., 2012. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. Proc. R. Soc. B 279, 3491–3500.
Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63, 1–19.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
Gatesy, J., Springer, M.S., 2013. Concatenation versus coalescence versus "concatalescence". Proc. Natl. Acad. Sci. USA 110, E1179.
Gatesy, J., Springer, M.S., 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. Mol. Phylogenet. Evol. 80, 231–266.
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.
Heled, J., Drummond, A.J., 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.
Hillis, D.M., Huelsenbeck, J.P., Cunningham, C.W., 1994. Application and accuracy of molecular phylogenies. Science 264, 671–677.
Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, NY, pp. 21–132.
Kubatko, L.S., Carstens, B.C., Knowles, L.L., 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25, 971–973.
Leaché, A.D., Rannala, B., 2011. The accuracy of species tree estimation under simulation: a comparison of methods. Syst. Biol. 60, 126–137.
Liu, L., 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24, 2542–2543.

Liu, L., Yu, L., 2010. Phybase: an R package for species tree analysis. Bioinformatics 26, 962–963.
Liu, L., Yu, L., 2011. Estimating species trees from unrooted gene trees. Syst. Biol. 60, 661–667.
Liu, L., Pearl, D.K., Brumfield, R.T., Edwards, S.V., 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62, 2080–2091.
Liu, L., Yu, L., Kubatko, L., Pearl, D.K., Edwards, S.V., 2009a. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53, 320–328.
Liu, L., Yu, L., Pearl, D.K., Edwards, S.V., 2009b. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. 58, 468–477.
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.
Liu, L., Xi, Z., Davis, C.C., 2015a. Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. Mol. Biol. Evol. 32, 791–805.
Liu, L., Xi, Z., Wu, S., Davis, C.C., Edwards, S.V., 2015b. Estimating phylogenetic trees from genome-scale data. Ann. N. Y. Acad. Sci. http://dx.doi.org/10.1111/nyas.12747.
McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754.
Meredith, R.W., Janečka, J.E., Gatesy, J., Ryder, O.A., Fisher, C.A., Teeling, E.C., Goodbla, A., Eizirik, E., Simão, T.L.L., Stadler, T., et al., 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334, 521–524.
Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30, i541–i548.
Mirarab, S., Bayzid, M.S., Warnow, T., 2015. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. http://dx.doi.org/10.1093/sysbio/syu1063.
Rambaut, A., Grassly, N.C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13, 235–238.
Rannala, B., Yang, Z., 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164, 1645–1656.
Rasmussen, M.D., Kellis, M., 2007. Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. Genome Res. 17, 1932–1942.
Reneker, J., Lyons, E., Conant, G.C., Pires, J.C., Freeling, M., Shyu, C.-R., Korkin, D., 2012. Long identical multispecies elements in plant and animal genomes. Proc. Natl. Acad. Sci. USA 109, E1183–E1191.
Seo, T.-K., 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25, 960–971.
Seo, T.-K., Kishino, H., Thorne, J.L., 2005. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. Proc. Natl. Acad. Sci. USA 102, 4436–4441.
Song, S., Liu, L., Edwards, S.V., Wu, S., 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA 109, 14942–14947.
Springer, M.S., Gatesy, J., 2014. Land plant origins and coalescence confusion. Trends Plant Sci. 19, 267–269.
Stamatakis, A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312–1313.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–771.
Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231.
William, J., Ballard, O., 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11, 334.
Wu, Y., 2012. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. Evolution 66, 763–775.
Wu, S., Song, S., Liu, L., Edwards, S.V., 2013. Reply to Gatesy and Springer: the multispecies coalescent model can effectively handle recombination and gene tree heterogeneity. Proc. Natl. Acad. Sci. USA 110, E1180.
Xi, Z., Rest, J.S., Davis, C.C., 2013. Phylogenomics and coalescent analyses resolve extant seed plant relationships. PLoS ONE 8, e80870.
Xi, Z., Liu, L., Rest, J.S., Davis, C.C., 2014. Coalescent versus concatenation methods and the placement of *Amborella* as sister to water lilies. Syst. Biol. 63, 919–932.
Zhong, B., Liu, L., Yan, Z., Penny, D., 2013. Origin of land plants using the multispecies coalescent model. Trends Plant Sci. 18, 492–495.
Zhong, B., Liu, L., Penny, D., 2014. The multispecies coalescent model and land plant origins: a reply to Springer and Gatesy. Trends Plant Sci. 19, 270–272.
Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion, The University of Texas at Austin.